National Research University Higher School of Economics

Manuscript

Dmitry Sergeevich Frolov

Aggregate Representation of Texts for Information Retrieval

Ph.D. Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science HSE

Moscow, 2019

This work was prepared at National Research University Higher School of Economics.

Academic Supervisor: Prof. Mirkin Boris Grigorevich, D.Sc., Tenured Professor,
National Research University Higher School of Economics

Dissertation Subject

Motivation

Problems of text analysis automation and information retrieval are becoming increasingly relevant due to modern processes of globalization and digitalization. Problems of text document search and retrieval are widely covered in modern literature, in particular, in [3, 16].

The relevance of this dissertation is determined by ever-increasing amounts of digitalized text data. Among the priorities there is a need for further progress in the following areas:

1) Increasing the performance of document search;
2) Improving the quality of document search;
3) Distributed storage of documents in the collection;
4) Parallel and non-simultaneous data processing;
5) Automation of text information analysis, including its structuring and interpreting.

Moving from sets of documents to their **aggregated representations** is one of the most efficient ways to advance in these areas. Aggregation, that is the process of combining, consolidation of certain objects on any common grounds to obtain generalized, aggregate indicators, can be considered as the transformation of the source data into a new model with significantly fewer variables or constraints which gives a different description of the studied process or object. Usually, aggregation methods use so-called feature representation of documents [2, 11], which is impossible without significant text pre-processing. There is another approach to aggregation that does not require preprocessing, in which arbitrary text fragments but not features are considered – the so-called annotated suffix tree (AST) method [7, 5]. Search methods involving AST can have certain advantages, for example, when performing document search on inaccurate queries (with spelling inaccuracies, errors). Methods based on the feature description are difficult to adapt to such situations.

Purpose and Objectives of the Research

The **purpose** of the dissertation research is to develop effective methods of text data search and analysis based on the use of annotated suffix trees to represent collections of documents, as well as software implementation of the developed methods, their experimental testing, and validation. To achieve this purpose, we are going:

1) To develop an efficient algorithm of text search based on representation by annotated suffix trees;
2) To develop versions of this algorithm for parallel and distributed computing, as well as non-simultaneous data processing;
3) To adapt the developed algorithms to work with dynamically changing collections of documents (cases of inserting, deleting, changing the text of documents);
4) To develop a methodology to interpret the results of information search using cluster analysis and taxonomic representation of the domain;
5) To develop mathematical software and to perform computational experiments to prove effectiveness of the developed programs;
6) To perform computational experiments to compare the development of the dissertation with the existing methods;
7) To apply the developed technology in the tasks of real information search and retrieval.

Object and Subject of the Research

The **object** of the dissertation research is the domain of information retrieval. The **subject** of the research is the use of aggregate text representation using AST for search problems in collections of text documents, collection structuring, interpreting of text document collections using taxonomies.

Research Methods and Reliability of the Results

The **methods** used in the research include:

1) A method of a text representation as an annotated suffix tree (AST) and a method of calculating string-to-document relevance based on AST;
2) Indexing and ranking methods for information retrieval problems;
3) Fuzzy clustering method FADDIS;
4) Taxonomic representation of domains, especially data science.

The **reliability** of the results is confirmed by the structure of mathematical computations, models and transformations used, software implementation testing, as well

as careful experimental validation of all developed methods and experiments in comparison with the proposed methods with existing approaches to similar problems.

The Results for the Thesis Defence

In the dissertation, the following main results are obtained:

1) A new method of information retrieval based on AST is developed (ASTS). A software implementation of the search system based on this method is carried out. Efficiency of the method is proved by computational experiments comparing ASTS with popular modern methods of search, including those specialized for fuzzy search problems. The comparison demonstrated the qualitative advantage of the developed method in fuzzy search objectives and a good balance of its quality characteristics and performance.
2) A method for interpretation of the retrieval results by using the structuring and optimal generalization of fuzzy clusters for a thematic set in taxonomy (ParGenFS) is developed. A software implementation of the developed method is made. The experimental approbation of the method ParGenFS is performed on the collections of papers' abstracts published by the Springer Journals on data science.
3) A method for the audience augmentation in Internet advertising (GUS) is developed. Advertising targeting is considered as a problem of information retrieval, the method is based on the optimal generalization of user segments in taxonomy of user interest segments. This application may be considered as an example of a problem in which the effect of optimal generalization is measurable. A new method of audience augmentation of advertising campaigns has been successfully implemented in practice.

Scientific Novelty of the Research

For the first time the method of information search for document collections represented in the form of annotated suffix trees (AST) is developed. A new method of information retrieval ASTS is developed. A software implementation of the search system based on this method is carried out. Method efficiency is proved by computational experiments comparing ASTS with popular modern methods of search, including those specialized for fuzzy search problems. The comparison showed the qualitative advantage of the developed method in fuzzy search objectives and a good balance of its quality characteristics and performance.

A method of interpreting the retrieval results by using the optimal generalization of fuzzy clusters of a thematic set in taxonomy (ParGenFS) is developed and implemented. Its experimental testing and validation is performed.

A method for the audience augmentation in Internet advertising is developed (GUS). Advertising targeting is considered as a problem of information retrieval. A software implementation and validation of the developed method to widen the audience of GUS using real advertising campaigns on the Internet is carried out. A new method of the audience augmentation of advertising campaigns has been successfully implemented in commercial company.

Theoretical Significance and Practical Value

The **theoretical significance** of the dissertation is that the dissertation introduces a novel method of information retrieval (ASTS) and its modifications for parallel, distributed, and dynamic computations, as well as the algorithm of optimal generalization of fuzzy clusters in taxonomies (ParGenFS). Further, ParGenFS method is applied for the following purposes:

1) Structuring and interpreting text collections;
2) Improving the effectiveness of advertising targeting, considered as a special case of information retrieval problem.

The **practical significance** of the thesis is supported by the following:

1) Effectiveness of ASTS method in the information retrieval problems (including fuzzy search) is experimentally proven;
2) Research Directions in Data Science are analyzed based on structuring and interpretation a collection of Springer Publications (1998–2017);
3) A method for effective audience augmentation of Internet advertising is offered and implemented.

Publicising and Publishing the Research Results

The research results were presented and discussed in the following conferences:

1) RuSSIR 2015 (Young Scientists Conference), poster presentation «Aggregate Text Representation for Information Retrieval in Collections of Text Documents», August, 24-28, 2015, Saint-Petersburg.

2) Summer School of Faculty of Computer Science, NRU HSE, poster presentation «Using Annotated Suffix Trees for Fuzzy Full Text Search», May, 27-29, 2016, Voronovo, Moskovskaya Obl.

3) RuSSIR-2016 (Young Scientists Conference), poster presentation «Using Annotated Suffix Trees for Fuzzy Full Text Search», August, 22-26, 2016, Saratov.

4) 3rd Kolmogorov's seminar on computational linguistics and language sciences, poster presentation «Annotation of a Document Collection by Finding Thematic Fuzzy Clusters and Parsimoniously Lifting Them in a Domain Taxonomy», April, 25, 2018, NRU HSE, Moscow.

5) All-Moscow seminar «Mathematical methods of decision analysis in economics, finance and politics», presentation «Annotation of a Document Collection by Finding Thematic Fuzzy Clusters and Parsimoniously Lifting Them in a Domain Taxonomy », May, 16, 2018, NRU HSE, Moscow.

6) All-Moscow seminar «Mathematical methods of decision analysis in economics, finance and politics», presentation «Generalization in taxonomies: model, method, applications», May, 15, 2019, NRU HSE, Moscow.

7) IARIA Content 2019, presentation «Method for Generalization of Fuzzy Sets», May, 5-9, 2019, Venice, Italy.

8) ICAISC-2019, poster presentation «Method for Generalization of Fuzzy Sets», June, 16-20, 2019, Zakopane, Poland.

9) IEEE 2019 International Conference on Fuzzy Systems, presentation «Using Taxonomy Tree to Generalize a Fuzzy Thematic Clusters», June, 23-26, 2019, New Orlean, USA.

10) World Congress on Global Optimization – 2019, presentation «Globally Optimal Parsimoniously Lifting a Fuzzy Query Set Over a Taxonomy Tree», July, 8-10, 2019, Metz, France.

11) IHIET-2019, poster presentation «A Method for Audience Extending in Programmatic Advertising by Using Parsimonious Generalization of User Segments», August, 22-24, 2019, Nice, France.

Higher-level publications:

1) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Using Taxonomy Tree to Generalize a Fuzzy Thematic Cluster // IEEE 2019 International Conference on Fuzzy Systems Proceedings, 2019. (CORE level A)

2) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Method for Generalization of Fuzzy Sets // Rutkowski L., Scherer R., Korytkowski M., Pedrycz W., Tadeusiewicz R., Zurada J. Artificial Intelligence and Soft Computing. ICAISC 2019. Lecture Notes in Computer Science, vol 11508. Springer, pp. 273-286. (Web of Science Q4, Scopus Q2)

Standard level publications:

3) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Globally Optimal Parsimoniously Lifting a Fuzzy Query Set Over a Taxonomy Tree // Le Thi H., Le H., Pham Dinh T. Optimization of Complex Systems: Theory, Models, Algorithms and Applications (WCGO). 2019. Advances in Intelligent Systems and Computing, vol 991. Springer, Cham, pp. 779-789. (Scopus Q3)

4) *Frolov D., Taran Z., Mirkin B.* A Method for Audience Extending in Programmatic Advertising by Using Parsimonious Generalization of User Segments // Human Interaction and Emerging Technologies (IHIET) 2019. (Scopus Q3)

5) *Frolov D.* Using Annotated Suffix Trees for Fuzzy Full Text Search // Communications in Computer and Information Science. Information Retrieval. 10th Russian Summer School, RuSSIR 2016. Revised Selected Papers. Springer. (Scopus Q3)

6) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Using Domain Taxonomy to Model Generalization of Thematic Fuzzy Clusters // IARIA Content 2019 Proceedings, pp. 20-25. (Web of Science)

7) *Frolov D.S.* Annotated suffix tree as a way of text representation for information retrieval in text collections // Business Informatics. 2015. No. 4 (34). P. 63–70. (Web of Science)

Other publications:

8) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Finding an appropriate generalization for a fuzzy thematic set in taxonomy // Working paper WP7/2018/04, Moscow, Higher School of Economics Publ. House, 2018, 60 P.

Length of the Dissertation

The dissertation consists of introduction, four chapters, conclusion and four appendices. The full dissertation volume is 189 pages, including 26 figures and 30 tables. The list of references contains 163 titles.

The Contents of the Dissertation

The **Introduction** describes the relevance of the dissertation theme, purpose, and objectives of the study, object and subject of the study, research methods description, main scientific results, theoretical and practical significance of the paper. The results of the dissertation, its approbation and the list of the author's publications on the research topic are presented. The structure of the dissertation is described.

The **first chapter** contains information retrieval problem formulation, necessary definitions, and main modern retrieval methods. The purpose of information search tasks is to provide users with a variety of documents that will meet their information needs. Therewith such needs should be formulated in an «understandable» form for the search mechanism. On the other hand, the entire set of data that is searched must also be presented in a format that allows the search mechanism to quickly identify potentially relevant documents. It is obvious that in both cases some information may be lost in the process of transformation. Therefore, organizing search mechanism algorithms and presenting information tasks are both important for search model development.

Several search models (Boolean, statistical and others) are described, their advantages and disadvantages are analyzed. This is followed by data presenting methods for information search tasks in collections of text documents. At present there is no single standard classification of approaches to text presentation but we can define the following main groups [3, 4]: symbolic models, pairwise superposition (alignment) of texts; different types of language models (formation of profiles and hidden Markov models, etc.); feature descriptions and models based on them; fragment presentations; vector representations (embeddings).

The dissertation describes the above methods, their advantages, and disadvantages. The emphasis is made on the aggregated presentation of texts using annotated suffix trees (AST). The advantage of this approach is that it can be applied without any pre-processing of these texts, unlike the feature approach which is impossible without pre-processing.

An annotated suffix tree (AST) — is a data structure used to compute and store all text fragments along with their frequencies. It is defined as a rooted tree in which each

node corresponds to one character and is marked with a frequency of the text fragment which encodes the path from the root to this node.

When using AST, the text is divided into strings - character chains. As a rule, one line is formed from 2-4 consecutive words. AST, which is built for text, allows to solve such a problem as a relevance calculation of a text string which will be used in the further work to solve the ranking problem.

Another important section of information retrieval is a fuzzy search (approximate search). It is a special type of search which allows an approximate match of query strings and text documents [1]. The practical significance of such tasks cannot be overestimated: documents in the natural language, as well as search queries, tend to have different kinds of inaccuracies such as omissions of letters, errors, misprints. There are many reasons why inaccuracies in the search query and documents in the collection may be formed when the spelling of words differs from their correct spelling: from misspelling by a person to the results of any text transformations. Nowadays due to the expansion of text recognition technologies (for example, from photos or scanned images), the text processing with inaccuracies due to such recognition is also relevant - the so-called OCR errors (Optical Character Recognition).

In recent years exploratory information retrieval (exploratory search) becomes more popular [18, 26]. This concept includes such tasks of a specific search when the user:

- Is not familiar with the domain of his interest (in other words the user needs to study this area to understand how to achieve the goal);
- Is not sure how to achieve his goal (in search technology);
- Can not clearly formulate the information need using the terms of available search technology.

It is obvious that the problem in this formulation is different from the classic one, when the search is performed on a specific search query. Exploratory search covers a broader class of activities in comparison with a typical information search and includes data structuring, analysis, interpretation, comparison of results. Therefore, combined strategies are often used to perform this type of search. It is quite important to choose an exploratory search method – to speed it up and obtain correct results.

The critical task of exploratory search is the interpretation of results. A lot of methods, such as collection structuring, give results which are sometimes extremely difficult to interpret. A classic example is the interpretation of the topics found in probabilistic topic

modeling, used, for example, to analyze hidden areas of research in the domain. A solution to this problem is the use of approaches based on domain representation in the form of taxonomies. One of the next chapters of the dissertation is devoted to the development of such approaches.

An important sub-task of exploratory search is the collection structuring: in exploratory search it is important to have methods of studying the collection document contents and collection structure (for example, semantic). For this purpose, clustering is the main group of methods [19, 21]. Clustering as a task of grouping the initial set into subsets-clusters of similar objects can show the internal structure of the collection, while the study of individual representatives of clusters can reveal their typical features. It is important to note that the existing methods of cluster description are based on the use of features with the same level of granularity (features, words) which were used to form clusters.

A separate set of exploratory search methods involves the use of taxonomies of the domain. Taxonomy, as a structure of concepts and a way to represent knowledge, can be used to structure and interpret search results or individual documents, and it also allows the concept of generalization, that is the transition to a large level of granularity.

The **second chapter** describes the developed method of information retrieval – ASTS, based on the aggregated text representation in AST form and inverted fragment indexing [6]. The results of experimental testing of this method on real data, as well as its comparison with other methods of information search, are presented: both in terms of quality metrics and performance. Experimental validation of the proposed method on real data is described. Several sets of experiments are presented in which the developed method was compared on collections from real data using user queries and on a specialized data set for information search. Both search quality and performance were compared.

The comparison involved ASTS search method (with a string length for AST of 3 consecutive words); search method based on BM25 ranking [3]; two search methods based on cosine vector similarity: of words and 3-character fragments; in both cases weighted with TF-IDF [3]; as well as methods based on PLSI and LDA (in *gensim* implementation, a modification using bigrams was also implemented for the latter). As far as test collections of documents were included into a document-oriented MongoDB database, a built-in full-text search mechanism was activated for the database. The experiment involved three collections: No. 1 – collection of documents based on the catalogue of the online store Ozon.ru; No. 2 – articles from web pages Habrahabr.ru; No. 3 – collection from a specialized TREC CAR dataset.

We present here the results of an experimental comparison of accuracy at the level of $N = 10$ documents in collection No. 1 for three groups of queries (1 – «Subcategory name» of the Ozone catalogue, 2 – «Explicit queries» and 3 – «Implicit queries») 1.

Table 1. Precision of the methods under consideration at the level $N = 10$ documents, Collection No. 1.

Number of a query group	ASTS	BM25	cos + TF-IDF words	cos + TF-IDF fragm.	PLSI	LDA	LDA with bigrams	Fulltext search MongoDB
1	0.83	0.86	0.79	0.76	0.70	0.85	0.86	0.51
2	0.82	0.84	0.72	0.70	0.68	0.81	0.86	0.52
3	0.77	0.43	0.44	0.65	0.41	0.43	0.55	0.21
Average	**0.81**	0.72	0.65	0.70	0.56	0.70	0.76	0.40

It should be noted that the derived accuracy estimates for ASTS method for the 3rd group of queries («Implicit queries») are almost the same as for the first two groups, while other methods, except for the approach based on the cosine fragment proximity weighted by TF-IDF, seriously lose. It is clear that this is due to the fact that ASTS method uses a fragment rather than a feature approach to text presentation. According to the results of quality and performance comparison ASTS method showed one of the best results among the methods under consideration.

Since one of the main ASTS advantages is the ability to perform fuzzy search, it was experimentally compared with special fuzzy search methods. ASTS method was compared with a popular algorithm based on calculation of Levenshtein distance (LD-based) [1] with n-gram reverse indexing, as well as with more modern methods of signature hashing (SH-based) [1] and Lemur search system. Computational experiments were carried out using real data. To conduct a fuzzy search two collections and special search queries were prepared. Collection No. 1 – a well-known standard collection «Reuters-21578» used for many tasks, including classical information search, No. 2 – collection of patent base USPTO, both – in English. All search queries were obtained from the strings of documents in the collection and sequences of their words by replacing the original words with versions containing errors.

The final comparison results of the considered four methods based on the experiment materials are given in the table 2.

As it can be seen from the table 2 ASTS method demonstrates the best balance of quality and performance among the methods under consideration.

Table 2. Results of comparison of fuzzy search methods (place 1-2 is assigned to a rank 1.5).

Metric	ASTS	LD-based	SH-based	Lemur
Search quality, place	1-2	1-2	3	4
Speed, place	3	4	2	1
Rank	**4.5**	5.5	5	5

As a result of the comparative experiments the main advantages of ASTS search method were revealed:

- It allows to avoid text preprocessing (lemmatization, stemming);
- It accepts texts with distorted features (errors, misprints in words);
- It demonstrates a good balance of search quality and performance.

The chapter also describes software implementation of ASTS method in a special search engine for Ubuntu operating systems with a console interface.

The **third chapter** describes optimal generalization algorithm for fuzzy sets (ParGenFS) and its application in exploratory search task.

As previously noted, exploratory search is an important and great problem of text collections analysis. A new approach to the problem using the domain taxonomy is presented on the example of generalization method for search results interpreting and later for demonstrating the overall strategy of document collection analysis.

The dissertation considers modeling of new research areas in the field of data science using a collection of research papers published by Springer Journals for 20 years and DST taxonomy [15] developed on the basis of computer science taxonomy by the Association for Computing Machinery (ACM) [28].

On the example of this collection and taxonomy, fuzzy clusters of taxonomy domains are found and analyzed in this research, each of which represents a development trend reflected in the collection. The main problem is the interpretation of such a fuzzy cluster. As already noted, interpretation of clusters is carried out with concepts of the same level of granularity (features, keywords) used in cluster construction. The dissertation offers to use the concept of coarsened granularity and taxonomy of the domain. In particular, we are talking about

generalization of taxonomy concepts corresponding to the cluster obtained. Here is a simple illustration of this approach in the following example (see figure 1).

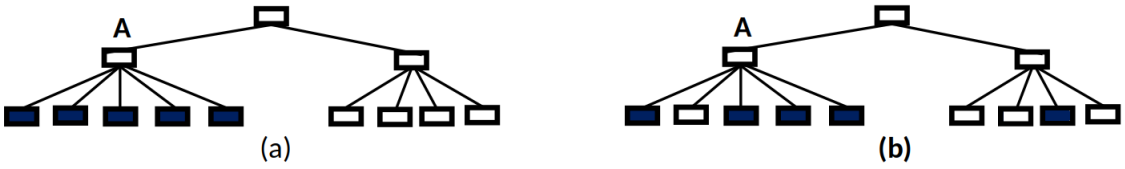

Fig. 1. Taxonomy fragment, black rectangles corresponding to elements of the cluster: simple case (a) and more complex (b).

Figure 1 (a) represents a taxonomy fragment with a leaf cluster fully covering all the children of the parental node A. Obviously, «lifting» of this cluster in the taxonomy to the vertex A is the most natural generalization of the cluster: it offers an interpretation of the cluster as all themes falling into the concept of A. The academic paper expands this approach to generalize less obvious cases – such as those presented in figure 1 (b).

The research presents a mathematical formalization of the generalization problem as of an optimal lifting of a fuzzy leaf cluster, defining a set, to higher ranks of the taxonomy and provides a recursive algorithm leading to a globally optimal solution to the problem. To demonstrate the task, we consider a hard set S shown with five black boxes in taxonomy leaves on a fragment of a tree in Figure (b). If we accept that a set S may be generalized by the root, this would lead to the situation when four white boxes are covered by the root and thus they fall in the same concept as S even as they do not belong to the set S. Such a situation will be referred to as a gap emergence, there are four gaps, and one head subject is introduced. Another lifting decision is lifting to the root of the left branch of the tree. We can see that the number of gaps has drastically decreased, to just 1. However, another oddity emerged: a black box on the right belonging to S but not covered by the root of the left branch at which the set S is mapped. This type of error will be referred to as an offshoot. At this lifting, three new items emerge: one head subject, one offshoot and one gap. To determine which lifting is preferable, we introduce penalty weights for the emergence of new items: 1 – penalty for a head node, λ – penalty for a gap γ – penalty for an offshoot. Then, for example, a penalty for the second lifting will be: $1 + \gamma + \lambda$. Based on the concepts introduced above this chapter defines a penalty function $p(H)$ for the set of head subjects H:

$$p(H) = \sum_{h \in H-I} u(h) + \sum_{h \in H-I} \sum_{g \in G(h)} \lambda v(g) + \sum_{h \in H \cap I} \gamma u(h). \quad (1)$$

Here I is a set of leaf nodes of the tree, $v(g)$ is the importance of the gap (it is logical that gaps can of different significance), $G(t)$ is a set of node gaps t, $V(t) = \sum_{g \in G(t)} v(g)$ is the total importance of gaps. This chapter proposes ParGenFS algorithm which allows to find optimal generalization of fuzzy thematic variety by minimizing this penalty function. Prior to algorithm use it is necessary to perform a tree pre-transformation; to do it the following steps should be taken:

1) All descendants of gaps should be removed from the tree.

2) To annotate all internal tree nodes with the membership function values based on the leaf membership value, for example, by the following rule: $u(h) = \sqrt{\sum_{t \in \chi(h)} u(t)^2}$ for $h \in T - I$, where $\chi(h)$ is a set of descendants of nodes h.

3) For each internal node t it is necessary to calculate sets of gaps $G(t)$ and the values of total gap importance $V(t) = \sum_{g \in G(t)} v(g)$ for later use in (1).

For each node of taxonomy tree t, algorithm ParGenFS calculates two sets – $H(t)$ and $L(t)$ containing such nodes in $T(t)$ in which there are acquisitions and losses of head nodes (including offshoots) respectively. The corresponding penalty is denoted $p(t)$. Algorithm outputs include values of the calculated sets in the root, in particular: H is a set of head nodes and offshoots, L is a set of gaps and p is an assigned penalty value. Algorithm pseudocode is given below.

ParGenFS algorithm

- **INPUT:** u, T
- **OUTPUT:** $H = H(root)$, $L = L(root)$, $p = p(root)$

I **Base case**

For each leaf $i \in I$

If $u(i) > 0$

$H(i) = \{i\}$, $L(i) = \oslash$, $p(i) = \gamma u(i)$

Else

$H(i) = \oslash$, $L(i) = \oslash$, $p(i) = 0$

II **Recursion**

If $u(t) + \lambda V(t) \leq \sum_{w \in \chi(t)} p(w)$

$H(t) = \{t\}$, $L(t) = G(t)$, $p(t) = u(t) + \lambda V(t)$

Else

$H(t) = \bigcup_{w \in \chi(t)} H(w)$, $L(t) = \bigcup_{w \in \chi(t)} L(w)$, $p(t) = \sum_{w \in \chi(t)} p(w)$

It is not too difficult to see that ParGenFS algorithm actually leads to an optimal lifting, as stated in the following theorem.

Theorem 1. *Any u-cover H found by the algorithm ParGenFS is a (global) minimizer of the penalty p (1).*

After consideration of the illustrative examples it is shown in the dissertation how ParGenFS can be applied to exploratory search task in the above collection of scientific publications. To do this the steps listed below are applied sequentially.

- Preparing a scholarly text collection;
- Preparing a taxonomy of the domain under consideration;
- Developing a matrix of relevance values between taxonomy leaf topics and research publications from the collection;
- Finding fuzzy clusters according to the matrix of relevance values;
- Lifting the clusters in the taxonomy to conceptualize them via generalization;
- Making conclusions from the generalizations.

After the first four steps, six clusters were obtained, three of which were particularly homogeneous. Based on their content, "Learning", "Retrieval" and "Clustering", we denote them L, R and C respectively. An example of a cluster ("Learning") is presented in the table 3.

All obtained clusters were lifted in the DST taxonomy using ParGenFS with the gap penalty $\lambda = 0.1$ and offshoot penalty $\gamma = 0.9$. The results of lifting cluster L are shown in the figure 2. The cluster has received three head subjects: machine learning, machine learning theory, and learning to rank. These represent the structure of the general concept "Learning" according to our text collection.

Table 3. Cluster L “Learning”: topics with the largest membership values.

$u(t)$	Code	Topic
0.300	5.2.3.8.	rule learning
0.282	5.2.2.1.	batch learning
0.276	5.2.1.1.2.	learning to rank
0.217	1.1.1.11.	query learning
0.216	5.2.1.3.3.	apprenticeship learning
0.213	1.1.1.10.	models of learning
0.203	5.2.1.3.5.	adversarial learning
0.202	1.1.1.14.	active learning

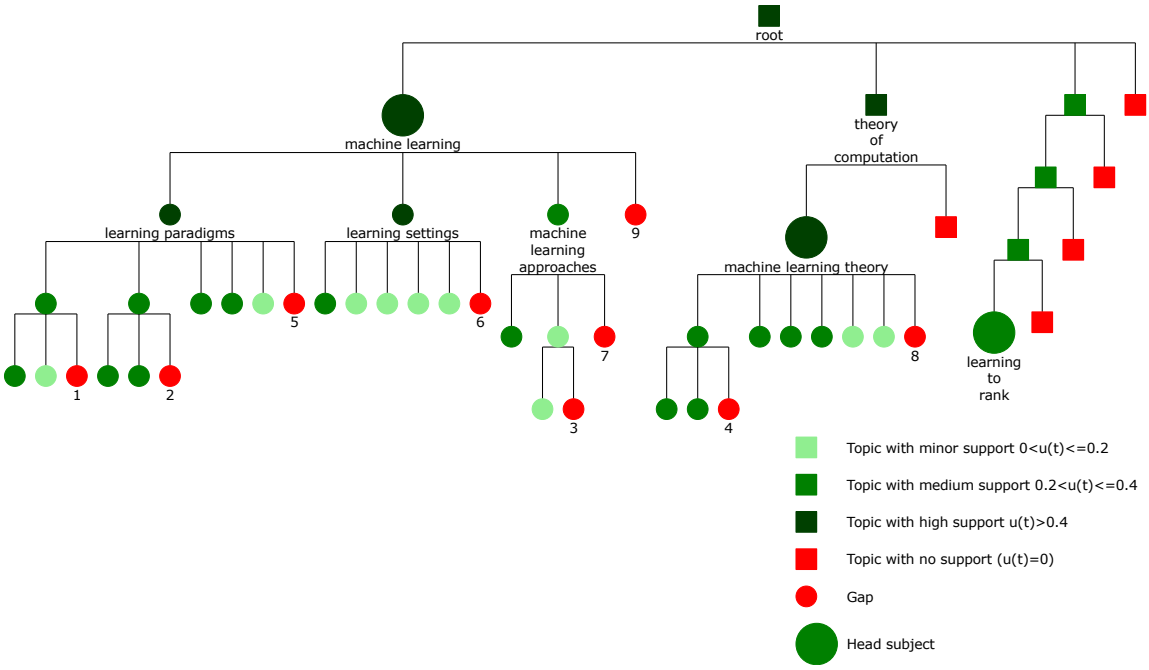

Fig. 2. Lifting results for the cluster L: Learning. Gaps are numbered.

Thematic clusters found in this collection of research papers form future development areas. In particular, it can be clearly seen from head subjects of the cluster “Learning” (see figure 2 and comments) that the main activity here is still focused on theory and methods rather than applications. Fields of machine learning, formerly focused mostly on tasks of learning subsets and partitions, are expanding currently towards learning of ranks and rankings.

The **fourth chapter** is devoted to the application of PanGenFS algorithm used to extend

the audience in advertising targeting (Programmatic). It considers a special applied case of information search problem: search for relevant audience among many users of an advertising campaign on the Internet. In this case, the search query is the requirements for the users who the advertisements should be displayed to, and the documents are the information about the users accumulated by the advertising system.

Recently the so-called programmatic approach [25] to Internet advertising is gaining popularity; it allows to show users in real time those advertising creatives which correspond to their interests revealed by the information accumulated in advertising systems about the user. In this approach the advertiser has an opportunity to purchase only the relevant audience. One of the targeting criteria is correspondence of user's interests to segment set for the advertising campaign. Each user, the advertising system has information about (for example, his browsing history), is associated with several segments of interests. Segments are elements of a taxonomy. Currently there are several taxonomies of user interests. One of the most popular taxonomies – IAB taxonomy [29] developed by the International Bureau of Internet Advertising, is used in the paper.

To increase the effectiveness of advertising campaigns it is necessary to solve the problem of augmenting the target audience: not always the number of users suitable for targeting is enough. One way of solving the problem is a look-a-like technique [24] when the audience is widening due to selection of users similar in some given metric to users from the nucleus (that is those users who come through targeting successfully). Another option is to weaken the boundaries of user's entry into the segments of interests. There is also an option of purchasing additional Internet traffic which results in additional financial costs for the advertising network.

This chapter offers another option to augment the target audience - by conceptual generalization of user segments. For this purpose, PanGenFS algorithm is applied which «lifts» user segments to the higher ranks of the taxonomy, whereby the user acquires head segments. The developed algorithm, the so-called generalization of user segments (GUS), deals with taxonomy of user interests, which can be represented by mentioned previously IAB taxonomy (or any other industrial taxonomy). This taxonomy covers traditional user interests presenting them in a form of a four-level rooted tree, nodes of which are labeled with topics of taxonomy. The «intellectual» widening of user audience occurs due to generalization: ParGenFS algorithm determines when generalization is possible based on values of user membership to segments. In its turn, targeting is performed with the participation of «extended» segments.

Our method was tested in three real advertising campaigns of the company «Natimatica» (`https://natimatica.com/`). The comparison involved three methods of advertising targeting: a classical method of segment programmatic targeting, a method which widens the target audience based on user segment generalization algorithm (GUS), a method which widens the target audience with a decrease in the thresholds of user membership to segments (DTS). According to the experiments results it may be concluded that application of user segments generalization algorithm to augment the audience of advertising campaigns allows to make larger amounts of ad impressions without significant drop in the audience quality. At the same time the algorithm based on lowering thresholds of user membership to segments, demonstrates audience augmenting with a noticeable decrease in its quality which was seen in the number of clicks on advertisements and CTR (ratio of clicks to ad impressions, click through rate). It should be noted that the effect of generalization can be measured in this application task.

As a measure of the effect it is natural to consider the ratio of the number of clicks according to OPS method to the number of clicks according to classic targeting (see the table 4).

Table 4. GUS method efficiency.

Advertising campaign	Classical targeting, number of clicks	GUS targeting, number of clicks	GUS method efficiency, %
1	1061	2544	239.8
2	201	367	182.6
3	749	1302	173.8

Obviously, the effect of optimal generalization method in the general task of exploratory search can be measured only when it is possible to formalize the process of knowledge level evaluation.

The main results obtained in the thesis and submitted for defence are listed in **conclusion**.

1) A new method of information retrieval (ASTS) is developed. A software implementation of the search system based on this method is carried out. Method efficiency is proved by computational experiments comparing ASTS with popular modern methods of search,

including those specialized for fuzzy search problems. The comparison demonstrated the qualitative advantage of the developed method in fuzzy search objectives and a good balance of its quality characteristics and performance.

2) A method of interpreting the search results and text collections by using the optimal generalization of fuzzy clusters for a thematic set in taxonomy (ParGenFS) is developed. A software implementation of the developed method is made. The experimental approbation of the method ParGenFS is performed on the set of abstracts published by the Springer Publishers in the field of data science.

3) A method for the efficient audience augmentation in Internet advertising (GUS) is developed. Advertising targeting is considered as a problem of information retrieval. A software implementation and validation of the developed method (using real advertising campaigns on the Internet) is carried out. A new method of the audience augmentation of advertising campaigns has been successfully implemented in a commercial company.

The **appendices** include: taxonomy of data science (on the whole) considered in the dissertation, some clusters obtained in the comparison of our method with the competitors, a listing of program code for AST implementation with an ability to calculate the relevance of a tree string, a listing of software implementation of ParGenFS algorithm, a listing of program code which displays a tree taxonomy with ParGenFS results.

References

[1] *Boytsov L. M.* Classification and experimental research of modern fuzzy dictionary search algorithms // Yandex. – 2004. – T. 6.

[2] *Korshunov A, Gomzin A.* Thematic Modeling of Texts in Natural Language // Trudy ISP RAN. 2012. №1. pp. 215-244.

[3] *Manning K. D., Raghavan P., Schuetze H.* Introduction to information retrieval / Manning K. D., Raghavan P., Schuetze H. - M .: Williams, 2011. – 680 p.

[4] *Mirkin B. G.* Cluster analysis methods for decision support: overview / B. G. Mirkin. - M .: Publishing House of the National Research University « Higher School of Economics», 2011 – 84 p.

[5] *Mirkin B. G., Chernyak E. L., Chugunova O. N.* An annotated suffix tree method for estimating the degree to which lines appear in text documents // Business Informatics. 2012. T. 3. № 21. pp. 31-41.

[6] *Frolov D.S.* Application of the annotated suffix tree method for search tasks in collections of text documents // Business Informatics. 2015. №. 4 (34). pp. 63–70.

[7] *Chernyak E. L., Mirkin B. G.* Using Internet resources to build taxonomy // In the book: Reports of the All-Russian Scientific Conference AIST 2013 / Otv. Ed .: E. L. Chernyak; scientific Ed .: D.I. Ignatov, M.Yu. Khachai, O. Barinova. M.: National Open University « INTUIT», 2013. pp. 36-48.

[8] *Altevogt P., Nitzsche R.* Method of generating a distributed text index for parallel query processing: patent 7966332 USA. – 2011.

[9] *Ashraf J., Chang E., Hussain O. K., Hussain F. K.* Ontology usage analysis in the ontology lifecycle: A state-of-the-art review // Knowledge-Based Systems, vol. 80, pp. 34-47, 2015.

[10] *Beneventano D., Dahlem N., El Haoum S., Hahn A., Montanari D., Reinelt M.* Ontology-driven semantic mapping // Enterprise Interoperability III, Part IV, Springer, pp. 329-341, 2008.

[11] *Blei D.* Probabilistic topic models // Communications of the ACM, 55 (4), pp. 77–84, 2012.

[12] *Chernyak E., Mirkin B.* Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources // Annals of Data Science, 2(1), pp. 61-82, 2015.

[13] *Frolov D.S.* Annotated suffix tree as a way of text representation for information retrieval in text collections //Business Informatics. 2015. No. 4 (34). pp. 63–70.

[14] *Frolov D.* Using Annotated Suffix Trees for Fuzzy Full Text Search, in: Communications in Computer and Information Science. Information Retrieval. 10th Russian Summer School, RuSSIR 2016, Saratov, Russia, August 22-26, 2016, Revised Selected Papers. Springer, 2016

[15] *Frolov D., Mirkin B., Nascimento S., Fenner T.* Finding an appropriate generalization for a fuzzy thematic set in taxonomy / Working paper WP7/2018/04, Moscow, Higher School of Economics Publ. House, 2018, 60 p.

[16] *Langville A. N., Meyer C. D.* Google PageRank and beyond: The science of search engine rankings // Princeton University Press, 2011.

[17] *Lloret E., Boldrini E., Vodolazova T., MartÃnez-Barco P., Munoz R., Palomar M.* A novel concept-level approach for ultra-concise opinion summarization // Expert Systems with Applications, 42(20), pp. 7148-7156, 2015.

[18] *Marchionini G.* Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41-46.

[19] *Mirkin B. G.* Core Concepts of Data Analysis / B. G. Mirkin. – Springer, 2012. – 416 p.

[20] *B. Mirkin, S. Nascimento, T. Fenner, L.M. Pereira* Building fuzzy thematic clusters and mapping them to higher ranks in a taxonomy, Int. Journal of Software Informatics, vol. 4, no. 3, pp. 257-275, 2010.

[21] *Mirkin B., Nascimento S.* Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices, Information Sciences, vol. 183, no. 1, pp. 16-34, 2012.

[22] *Mueller G., Bergmann R.* Generalization of Workflows in Process-Oriented Case-Based Reasoning // FLAIRS Conference, pp. 391-396, 2015.

[23] *Nascimento S., Fenner T., Mirkin B.* Representing research activities in a hierarchical ontology // Procs. of 3rd International Workshop on Combinations of Intelligent Methods and Applications (CIMA 2012), Montpellier, France, August, 28, pp. 23-29, 2012.

[24] *Popov A., Iakovleva D.* Adaptive look-alike targeting in social networks advertising // Procedia Computer Science. – 2018. – T. 136. – pp. 255-264.

[25] *Sayedi A.* Real-Time Bidding in Online Display Advertising, 2017.

[26] *White R., Roth R.* Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

[27] *Yuan, Y., Wang, F., Li, J., Qin, R.* A survey on real time bidding advertising. In Service Operations and Logistics, and Informatics (SOLI) // 2014 IEEE International Conference. IEEE. pp. 418-423.

[28] The 2012 ACM Computing Classification System [Electronic resource]. 2019 –. – Access: *http://www.acm.org* /about/class/2012 , free available.

[29] IAB Tech Lab Content Taxonomy [Electronic resource]. 2019 –. – Access: *https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag- taxonomy/*, free available.

[30] OpenRTB Protocol [Electronic resource]. 2019 –. – Access: *https://www.iab.com/guidelines/real-time-bidding-rtb-project/* , free available.