Using Domain Taxonomy to Model Generalization of Thematic Fuzzy Clusters

Dmitry Frolov
National Research University "Higher School of Economics" Moscow, Russian Federation
Email: dfrolov@hse.ru

Susana Nascimento
Universidade Nova de Lisboa Caparica, Portugal
Email: snt@fct.unl.pt

Trevor Fenner
Birkbeck, University of London London, UK
Email: trevor@dcs.bbk.ac.uk

Boris Mirkin
National Research University "Higher School of Economics" Moscow, Russian Federation, and Birkbeck, University of London London, UK
Email: bmirkin@hse.ru

***Abstract*—We define a most specific generalization of a fuzzy set of topics assigned to leaves of the rooted tree of a domain taxonomy. This generalization lifts the set to its "head subject" in the higher ranks of the taxonomy tree. The head subject is supposed to "tightly" cover the query set, possibly bringing in some errors, both "gaps" and "offshoots". Our method globally minimizes a penalty function combining the numbers of head subjects and gaps and offshoots, differently weighted. We apply this to a collection of about 18000 research papers published in Springer journals on Data Science for the past 20 years. We extract a taxonomy of Data Science from the international Association for Computing Machinery Computing Classification System 2012 (ACM-CCS). We find fuzzy clusters of leaf topics over the text collection and use lifted head subjects of the thematic clusters to comment on the tendencies of current research in the corresponding aspects of the domain.**

Keywords–Generalization; gap-offshoot penalty; fuzzy cluster; spectral clustering; annotated suffix tree.

I. INTRODUCTION

The issue of automation of structurization and interpretation of digital text collections is of ever-growing importance because of both practical needs and theoretical necessity. This paper is concerned with an aspect of this, modeling generalization as a unique feature of human cognitive abilities.

The existing approaches to computational analysis of structure of text collections usually involve no generalization as a specific aim. The most popular tools for structuring text collections are cluster analysis and topic modelling. Both involve items of the same level of granularity as individual words or short phrases in the texts, thus no generalization as an explicitly stated goal.

Nevertheless, the hierarchical nature of the universe of meanings is reflected in the flow of publications on text analysis. We can distinguish between at least three directions at which the matter of generalization is addressed. First of all, there are activities related to developing taxonomies, especially those involving hyponymic/hypernymic relations (see, for example, [14] [17], and references therein). A recent paper [15] is devoted to supplementing a taxonomy with newly emerging research topics.

Another direction is part of conventional activities in text summarization. Usually, summaries are created using a rather mechanistic approach of sentence extraction. There is, however, also an approach for building summaries as abstractions of texts by combining some templates, such as Subject-Verb-Object (SVO) triplets (see, for example, [9]).

One more direction is what can be referred to as "operational" generalization. In this direction, the authors use generalized case descriptions involving taxonomic relations between generalized states and their parts to achieve a tangible goal, such as improving characteristics of text retrieval (see, for example, [12] [16].)

This paper begins a novel direction of research by using an existing taxonomy for straightforwardly implementing the idea of generalization. According to the Merriam-Webster dictionary, the term "generalization" refers to deriving a general conception from particulars. The "particulars", in our case, are represented by a fuzzy set of taxonomy leaves, whereas "the general conception" will be represented by a higher rank taxonomy node to embrace the fuzzy set as tight as possible. To the best of our knowledge, this approach has been never explored before. We experimentally show that our method leads to the type of conclusions which cannot be provided by other existing approaches to the analysis of text collections (see the end of Section III).

Our text collection is a set of about 18,000 research papers published by the Springer Publishers in 17 journals related to Data Science for the past 20 years. Our taxonomy of Data Science is a slightly modified part of the world-wide Association for Computing Machinery Computing Classification System (ACM-CCS), a 5-layer taxonomy published in 2012 [1].

The rest of the paper is organized accordingly. Section II presents a mathematical formalization of the generalization problem as of parsimoniously lifting of a given fuzzy leaf set to higher ranks of the taxonomy and provides a recursive algorithm leading to a globally optimal solution to the problem. Section III describes an application of this approach to deriving tendencies in development of the Data Science according to our Springer text collection mapped to the ACM-CCS. Its subsections describe stages of our approach to finding and generalizing fuzzy clusters of research topics. In the end, we point to tendencies in the development of the corresponding parts of Data Science, as drawn from the generalization results.

II. GENERALIZATION BY PARSIMONIOUSLY LIFTING A FUZZY THEMATIC SUBSET IN TAXONOMY: MODEL AND METHOD

Mathematically, a taxonomy is a rooted tree whose nodes are annotated by taxonomy topics.

We consider the following problem. Given a fuzzy set S of taxonomy leaves, find a node $t(S)$ of higher rank in the taxonomy, that covers the set S in a most specific way. Such a "lifting" problem is a mathematical explication of the human facility for generalization.

 ISBN: 978-1-61208-707-8

The problem is not as simple as it may seem to be. Consider, for the sake of simplicity, a hard set S shown with five black leaf boxes on a fragment of a tree in Figure 1. Figure 2 illustrates the situation at which the set of black boxes is lifted to the root, which is shown by blackening the root box, and its offspring, too. If we accept that set S may be generalized by the root, this would lead to a number, four, white boxes to be covered by the root and, thus, in this way, falling in the same concept as S even as they do not belong in S. Such a situation will be referred to as a gap. Lifting with gaps should be penalized. Altogether, the number of conceptual elements introduced to generalize S here is 1 head subject, that is, the root to which we have assigned S, and the 4 gaps occurred just because of the topology of the tree, which imposes this penalty. Another lifting decision is illustrated in Figure 3: here the set is lifted just 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 belongs to S but is not covered by the head subject in the root of the left branch. 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. Which of the errors is greater?

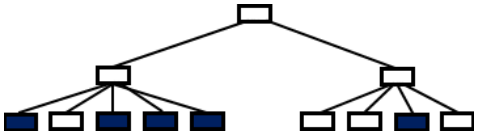

Figure 1. A crisp query set, shown by black boxes, to be conceptualized in the taxonomy.

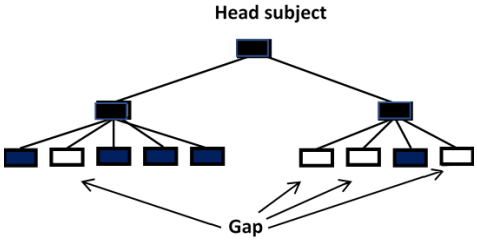

Figure 2. Generalization of the query set from Figure 1 by mapping it to the root, with the price of four gaps emerged at the lift.

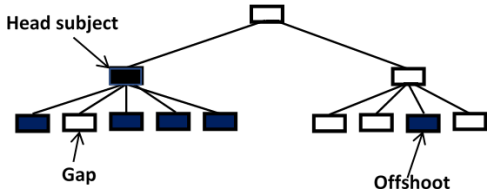

Figure 3. Generalization of the query set from Figure 1 by mapping it to the root of the left branch, with the price of one gap and one offshoot emerged at this lift.

We are interested to see whether a fuzzy set S can be generalized by a node t from higher ranks of the taxonomy, so that S can be thought of as falling within the framework covered by the node t. The goal of finding an interpretable pigeon-hole for S within the taxonomy can be formalized as that of finding one or more "head subjects" t to cover S with the minimum number of all the elements introduced at the generalization: head subjects, gaps, and offshoots. This goal realizes the principle of Maximum Parsimony (MP).

Consider a rooted tree T representing a hierarchical taxonomy so that its nodes are annotated with key phrases signifying various concepts. We denote the set of all its *leaves* by I. The relationship between nodes in the hierarchy is conventionally expressed using genealogical terms: each node $t \in T$ is said to be the *parent* of the nodes immediately descending from t in T, its *children*. We use $\chi(t)$ to denote the set of children of t. Each *interior* node $t \in T - I$ is assumed to correspond to a concept that generalizes the topics corresponding to the leaves $I(t)$ descending from t, viz. the leaves of the subtree $T(t)$ rooted at t, which is conventionally referred to as the *leaf cluster of* t.

A *fuzzy set* on I is a mapping u of I to the non-negative real numbers that assigns a membership value, or support, $u(i) \geq 0$ to each $i \in I$. We refer to the set $S_u \subset I$, where $S_u = \{i \in I : u(i) > 0\}$, as the *base* of u. In general, no other assumptions are made about the function u, other than, for convenience, commonly limiting it to not exceed unity. Conventional, or *crisp*, sets correspond to binary membership functions u such that $u(i) = 1$ if $i \in S_u$ and $u(i) = 0$ otherwise.

Given a fuzzy set u defined on the leaves I of the tree T, one can consider u to be a (possibly noisy) projection of a general concept, u's "head subject", onto the corresponding leaf cluster. Under this assumption, there should exist a head subject node h among the interior nodes of the tree T such that its leaf cluster $I(h)$ more or less coincides (up to small errors) with S_u. This head subject is the generalization of u to be found. The two types of possible errors associated with the head subject, if it does not cover the base precisely, are false positives and false negatives, referred to in this paper, as *gaps* and *offshoots*, respectively. They are illustrated in Figures 2 and 3. Given a head subject node h, a gap is a node t covered by h but not belonging to u, so that $u(t) = 0$. In contrast, an offshoot is a node t belonging to u so that $u(t) > 0$ but not covered by h. Altogether, the total number of head subjects, gaps, and offshoots has to be as small as possible. To this end, we introduce a penalty for each of these elements. Assuming for the sake of simplicity, that the black box leaves on Figure 1 have membership function values equal to unity, one can easily see that the total penalty at the head subject raised to the root (Figure 2) is equal to $1+4\lambda$ where 1 is the penalty for a head subject and λ, the penalty for a gap, since the lift on Figure 2 involves one head subject, the root, and four gaps, the blank box leaves. Similarly, the penalty for the lift on Figure 3 to the root of the left-side subtree is equal to $1 + \gamma + \lambda$ where γ is the penalty for an offshoot, as there is one copy of each, head subject, gap, and offshoot, in Figure 3. Therefore, depending on the relationship between γ and λ either lift on Figure 2 or lift on Figure 3 is to be chosen.

Consider a candidate node h in T and its meaning relative to fuzzy set u. An *h-gap* is a node g of $T(h)$, other than h, at which a *loss* of the meaning has occurred, that is, g is a maximal u-irrelevant node in the sense that its parent is not u-irrelevant. Conversely, establishing a node h as a head subject can be considered as a *gain* of the meaning of u at the node. The set of all h-gaps will be denoted by $G(h)$. A node $t \in T$ is referred to as *u-irrelevant* if its leaf-cluster $I(t)$ is disjoint from the base S_u. Obviously, if a node is u-irrelevant, all of

its descendants are also u-irrelevant.

An *h-offshoot* is a leaf $i \in S_u$ which is not covered by h, i.e., $i \notin I(h)$. The set of all h-offshoots is $S_u - I(h)$. Given a fuzzy topic set u over I, a set of nodes H will be referred to as a *u-cover* if: (a) H covers S_u, that is, $S_u \subseteq \bigcup_{h \in H} I(h)$, and (b) the nodes in H are unrelated, i.e., $I(h) \cap I(h') = \emptyset$ for all $h, h' \in H$ such that $h \neq h'$. The interior nodes of H will be referred to as *head subjects* and the leaf nodes as *offshoots*, so the set of offshoots in H is $H \cap I$. The set of *gaps* in H is the union of $G(h)$ over all head subjects $h \in H - I$.

We define the penalty function $p(H)$ for a u-cover H as:

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

The problem we address is to find a u-cover H that globally minimizes the penalty $p(H)$. Such a u-cover is the parsimonious generalization of the set u.

Before applying an algorithm to minimize the total penalty, one needs to execute a preliminary transformation of the tree by pruning it from all the non-maximal u-irrelevant nodes, i.e., descendants of gaps. Simultaneously, the sets of gaps $G(t)$ and the internal summary gap importance $V(t) = \sum_{g \in G(t)} v(g)$ in (1) can be computed for each interior node t. We note that the elements of S_u are in the leaf set of the pruned tree, and the other leaves of the pruned tree are precisely the gaps. After this, our lifting algorithm ParGenFS applies. For each node t, the algorithm ParGenFS computes two sets, $H(t)$ and $L(t)$, containing those nodes in $T(t)$ at which respectively gains and losses of head subjects occur (including offshoots). The associated penalty $p(t)$ is computed too.

An assumption of the algorithm is that no gain can happen after a loss. Therefore, $H(t)$ and $L(t)$ are defined assuming that the head subject has not been gained (nor therefore lost) at any of t's ancestors. The algorithm ParGenFS recursively computes $H(t)$, $L(t)$ and $p(t)$ from the corresponding values for the child nodes in $\chi(t)$.

Specifically, for each leaf node that is not in S_u, we set both $L(\cdot)$ and $H(\cdot)$ to be empty and the penalty to be zero. For each leaf node that is in S_u, $L(\cdot)$ is set to be empty, whereas $H(\cdot)$, to contain just the leaf node, and the penalty is defined as its membership value multiplied by the offshoot penalty weight γ. To compute $L(t)$ and $H(t)$ for any interior node t, we analyze two possible cases: (a) when the head subject has been gained at t and (b) when the head subject has not been gained at t.

In case (a), the sets $H(\cdot)$ and $L(\cdot)$ at its children are not needed. In this case, $H(t)$, $L(t)$ and $p(t)$ are defined by:

$$H(t) = \{t\}, \quad L(t) = G(t), \quad p(t) = u(t) + \lambda V(t). \quad (2)$$

In case (b), the sets $H(t)$ and $L(t)$ are just the unions of those of its children, and $p(t)$ is the sum of their penalties:

$$H(t) = \bigcup_{w \in \chi(t)} H(w), \quad L(t) = \bigcup_{w \in \chi(t)} L(w), \\ p(t) = \sum_{w \in \chi(t)} p(w). \quad (3)$$

To obtain a parsimonious lift, whichever case gives the smaller value of $p(t)$ is chosen.

When both cases give the same values for $p(t)$, we may choose, say, (a). The output of the algorithm consists of the values at the root, namely, H – the set of head subjects and offshoots, L – the set of gaps, and p – the associated penalty.

It was mathematically proven that the algorithm ParGenFS leads to an optimal lifting indeed [5].

III. Highlighting tendencies in the current research by clustering and lifting a collection of research papers

Being confronted with the problem of structuring and interpreting a set of research publications in a domain, one can think of either of the following two pathways to take. One is so-to-speak empirical and the other theoretical. The first pathway tries to discern main categories from the texts, the other, from knowledge of the domain. The first approach is exemplified by the LDA-based topic modeling [2]; the second approach, by using an expert-driven taxonomy, such as ACM-CCS [1] (see, for example, [13]).

This paper follows the second pathway by moving, in sequence, through the following stages:

- 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 structure of relevance values;
- lifting the clusters over the taxonomy to conceptualize them via generalization;
- making conclusions from the generalizations.

Each of the items is covered in a separate subsection further on.

A. Scholarly text collection

Because of a generous offer from the Springer Publisher, we were able to download a collection of 17685 research papers together with their abstracts published in 17 journals related to Data Science for 20 years from 1998-2017 [5]. We take the abstracts to these papers as a representative collection.

B. DST Taxonomy

Taxonomy building is a form of knowledge engineering which is getting more and more popular. Most known are taxonomies within the bioinformatics Genome Ontology (GO) project [6], Health and Medicine SNOMED CT project [8] and the like. Mathematically, a taxonomy is a rooted tree, a hierarchy, whose all nodes are labeled by main concepts of the domain the taxonomy relates to. The hierarchy corresponds to the inclusion relation: the fact that node A is the parent of B means that B is part, or a special case, of A.

The subdomain of our choice is Data Science, comprising such areas as machine learning, data mining, data analysis, etc. We take that part of the ACM-CCS 2012 taxonomy, which is related to Data Science, and add a few leaves related to more recent Data Science developments. The Taxonomy of Data Science, DST, with all its 317 leaves, is presented in [5].

C. Deriving fuzzy clusters of taxonomy topics

Clusters of topics should reflect co-occurrence of topics: the greater the number of texts to which both t and t' topics are relevant, the greater the interrelation between t and t',

the greater the chance for topics t and t' to fall in the same cluster. We have tried several popular clustering algorithms at our data. Unfortunately, no satisfactory results have been found. Therefore, we present here results obtained with the FADDIS algorithm developed in [11] specifically for finding thematic clusters. This algorithm implements assumptions that are relevant to the task:

- LN Laplacian Pseudo-Inverse Normalization (LaPIN): Similarity data transformation, modeling – to an extent – heat distribution and, in this way, making the cluster structure sharper.
- AA Additivity: Thematic clusters behind the texts are additive, so that co-relevance similarity values are sums of contributions by different hidden themes.
- AN Non-Completeness: Clusters do not necessarily cover all the key phrases available, as the text collection under consideration may be irrelevant to some of them.

1) Co-relevance topic-to-topic similarity score: Given a keyphrase-to-document matrix R of relevance scores is converted to a keyphrase-to-keyphrase similarity matrix A for scoring the "co-relevance" of keyphrases according to the text collection structure. The similarity score $a_{tt'}$ between topics t and t' is computed as the inner product of vectors of scores $r_t = (r_{tv})$ and $r_{t'} = (r_{t'v})$ where $v = 1, 2, \ldots, V = 17685$. The inner product is moderated by a natural weighting factor assigned to texts in the collection. The weight of text v is defined as the ratio of the number of topics n_v relevant to it and n_{max}, the maximum n_v over all v = 1,2,...,V. A topic is considered relevant to v if its relevance score is greater than 0.2 (a threshold found experimentally, see [4]).

2) Fuzzy thematic clusters: To obtain fuzzy clusters of topics we used a method FADDIS, that was developed in [10]. FADDIS finds clusters one-by-one. Paper [11] provides some theoretical and experimental computation results to demonstrate that FADDIS is competitive over popular fuzzy clustering approaches.

After computing the 317×317 topic-to-topic co-relevance matrix, converting it to a topic-to-topic LaPIN transformed similarity matrix, and applying FADDIS clustering, we sequentially obtained 6 clusters, of which three clusters appear to be obviously homogeneous. They relate to "Learning", "Retrieval", and "Clustering". These clusters, L, R, and C, are presented in Tables I, II, and III, respectively.

TABLE I. CLUSTER L "LEARNING": TOPICS WITH MEMBERSHIP VALUES GREATER THAN 0.15

$u(t)$	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
0.192	5.2.1.4.1.	Transfer Learning
0.192	5.2.1.4.2.	Lifelong Machine learning
0.189	1.1.1.8.	Online Learning Theory
0.166	5.2.2.2.	Online Learning Settings
0.159	1.1.1.3.	Unsupervised Learning and Clustering

TABLE II. CLUSTER R "RETRIEVAL": TOPICS WITH MEMBERSHIP VALUES GREATER THAN 0.15

$u(t)$	Code	Topic
0.211	3.4.2.1.	Query Representation
0.207	5.1.3.2.1.	Image Representations
0.194	5.1.3.2.2.	Shape Representations
0.194	5.2.3.6.2.1	Tensor Representation
0.191	5.2.3.3.3.2	Fuzzy Representation
0.187	3.1.1.5.3.	Data Provenance
0.173	2.1.1.5.	Equational Models
0.173	3.4.6.5.	Presentation of Retrieval Results
0.165	5.1.3.1.3.	Video Segmentation
0.155	5.1.3.1.2.	Image Segmentation
0.154	3.4.5.5.	Sentiment Analysis

TABLE III. CLUSTER C "CLUSTERING": TOPICS WITH MEMBERSHIP VALUES GREATER THAN 0.15

$u(t)$	Code	Topic
0.327	3.2.1.4.7	Biclustering
0.286	3.2.1.4.3	Fuzzy Clustering
0.248	3.2.1.4.2	Consensus Clustering
0.220	3.2.1.4.6	Conceptual Clustering
0.192	5.2.4.3.1	Spectral Clustering
0.187	3.2.1.4.1	Massive Data Clustering
0.159	3.2.1.7.3	Graph Based Conceptual Clustering
0.151	3.2.1.9.2.	Trajectory Clustering

D. Results of lifting clusters L, R, and C within DST

To apply ParGenFS algorithm, values of λ and γ should be defined first. This may highly affect the results. In the example above, lifting in Figure 2 is more parsimonious than lifting in Figure 3 if $\gamma > 3\lambda$, or the latter, if otherwise. We define off-shoot penalty $\gamma = 0.9$ to make it almost as costly as a head subject. In contrast, the gap penalty is defined as $\lambda = 0.1$ to take into account that every node in the taxonomy tree has about 10-15 children so that half-a-dozen gaps would be admissible. The clusters above are lifted in the DST taxonomy using ParGenFS algorithm with these parameter values.

The results of lifting of Cluster L are shown in Figure 4. There are three head subjects: machine learning, machine learning theory, and learning to rank. These represent the structure of the general concept "Learning" according to the text collection under consideration. The list of gaps obtained is less instructive, reflecting probably a relatively modest coverage of the domain by the publications in the collection (see in Table IV).

Similar comments can be made with respect to results of lifting of Cluster R: Retrieval. The obtained head subjects: Information Systems and Computer Vision show the structure of "Retrieval" in the set of publications under considerations.

For Cluster C 16 (!) head subjects were obtained: clustering, graph based conceptual clustering, trajectory clustering, clustering and classification, unsupervised learning and clustering, spectral methods, document filtering, language models, music retrieval, collaborative search, database views, stream management, database recovery, mapreduce languages, logic and databases, language resources. As one can see, the core clustering subjects are supplemented by methods and environments in the cluster – this shows that the ever increasing role of clustering activities perhaps should be better reflected in the taxonomy.

 ISBN: 978-1-61208-707-8

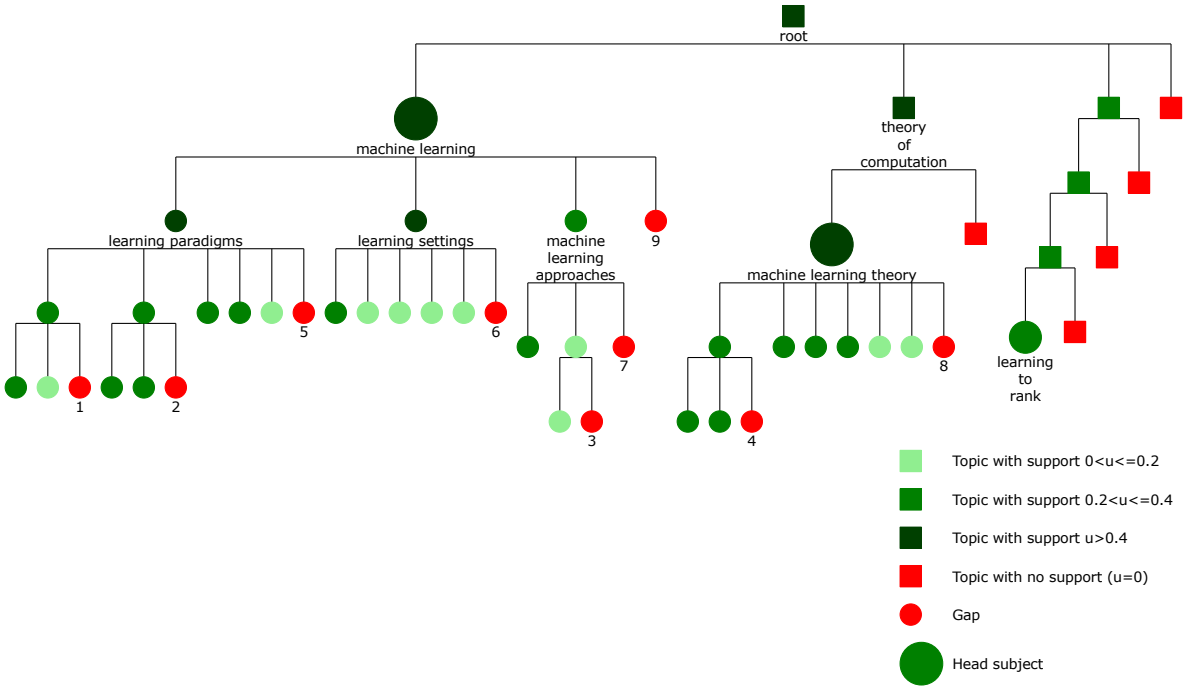

Figure 4. Lifting results for Cluster L: Learning. Gaps are numbered, see Table IV.

TABLE IV. GAPS AT THE LIFTING OF CLUSTER L

Number	Topics
1	ranking, supervised learning by classification, structured outputs
2	sequential decision making in practice, inverse reinforcement learning in practice
3	statistical relational learning
4	sequential decision making, inverse reinforcement learning
5	unsupervised learning
6	learning from demonstrations, kernel approach
7	classification and regression trees, kernel methods, neural networks, learning in probabilistic graphical models, learning linear models, factorization methods, markov decision processes, stochastic games, learning latent representations, multiresolution, support vector machines
8	sample complexity and generalization bounds, boolean function learning, kernel methods, boosting, bayesian analysis, inductive inference, structured prediction, markov decision processes, regret bounds
9	machine learning algorithms

E. Making conclusions

We can see that the topic clusters found with the text collection do highlight areas of soon-to-be developments. Three clusters under consideration closely relate, in respect, to the following processes:

- theoretical and methodical research in learning, as well as merging the subject of learning to rank within the mainstream;
- representation of various types of data for information retrieval, and merging that with visual data and their semantics; and
- various types of clustering in different branches of the taxonomy related to various applications and instruments.

In particular, one can see from the “Learning” head subjects (see Figure 4 and comments to it) that main work here still concentrates on theory and method rather than applications. A good news is that the field of learning, formerly focused mostly on tasks of learning subsets and partitions, is expanding currently towards learning of ranks and rankings. Of course, there remain many sub-areas to be covered: these can be seen in and around the list of gaps in Table IV.

Moving to the lifting results for the information retrieval cluster R, we can clearly see the tendencies of the contemporary stage of the process. Rather than relating the term “information” to texts only, as it was in the previous stages of the process of digitalization, visuals are becoming parts of the concept of information. There is a catch, however. Unlike the multilevel granularity of meanings in texts, developed during millennia of the process of communication via languages in the humankind, there is no comparable hierarchy of meanings for images. One may only guess that the elements of the R cluster related to segmentation of images and videos, as well as those related to data management systems, are those that are going to be put in the base of a future multilevel system of meanings for images and videos. This is a direction for future developments clearly seen from lifting results.

Regarding the “clustering” cluster C with its 16 (!) head subjects, one may conclude that, perhaps, a time moment has come or is to come real soon, when the subject of clustering must be raised to a higher level in the taxonomy to embrace all these “heads”. At the beginning of the Data Science era, a few decades ago, clustering was usually considered a more-or-less

auxiliary part of machine learning, the unsupervised learning. Perhaps, soon we are going to see a new taxonomy of Data Science, in which clustering is not just an auxiliary instrument but rather a model of empirical classification, a big part of the knowledge engineering.

It should be pointed out that analysis of tendencies of research is carried out by several groups using co-citation data, especially in dynamics (see, for example, a review in [3]). This approach leads to conclusions involving "typical", rather than general, authors and/or papers, and, therefore, is complementary to our approach.

IV. Conclusion and Future Work

This paper presents a formalization of the concept of generalization, an important part of the human ability for conceptualization. According to Collins Dictionary, conceptualization is "formation (of a concept or concepts) out of observations, experience, data, etc." We assume that such an operation may require a coarser granularity of the domain structuring. This is captured by the idea of lifting a query set to higher ranks in a hierarchical taxonomy of the domain.

The hierarchical structure of taxonomy brings in possible inconsistencies between a query set and the taxonomy structure. These inconsistencies can be of either of two types, gaps or offshoots, potentially emerging at the coarser "head subject" to cover the query set. A gap is such a node of the taxonomy, that is covered by the head subject but does not belong in the query set. An offshoot is a node of the taxonomy, that does belong in the query set but is not covered by the head subject. The higher the rank of a candidate for the conceptual head subject, the larger the number of gaps. The lower is the rank of the head subject, the larger the number of offshoots. Our algorithm ParGenFS allows to find a globally optimal lifting to balance the numbers of head subjects, gaps, and offshoots depending on relative penalties for these types of inconsistencies.

The proposed approach to generalization can be used in a number of similar tasks, such as positioning of a research project, interpretation of a concept which is not present in the taxonomy, annotation of a set of research articles. These all are parts of the processes of long-term research analysis and planning at which our approach should be positioned.

Among major issues requiring further development in this direction, two of the most relevant are taxonomy developments and specifying penalty weights. The former needs more attention both from research communities and planning committees. Specifically, most urgent directions for development here are: developing better methods to automate the process of taxonomy making and open discussion of the taxonomies at conferences and meetings of research communities and committees. Our current approach could be used for automation of updating taxonomies at the situations at which there are too many head subjects, like in the case of "Clustering" cluster in this paper. As to the latter, a reasonable computational progress over penalty weights can be achieved, in our view, by replacing the criterion of maximum parsimony by the criterion of maximum likelihood if each node of the taxonomy can be assigned probabilities of "gain" and "loss" of topic events.

Acknowledgment

D.F. and B.M. acknowledge continuing support by the Academic Fund Program at the National Research University Higher School of Economics (grant 19-04-019 in 2018-2019) and by the International Decision Choice and Analysis Laboratory (DECAN) NRU HSE, in the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the the Russian Academic Excellence Project "5-100". S. N. acknowledges the support by FCT/MCTES, NOVA LINCS (UID/CEC/04516/2013)

References

[1] The 2012 ACM Computing Classification System. [Online]. Available:*http://www.acm.org* /about/class/2012 (Retrieved 17 March, 2019).

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

[3] C. Chen, "Science mapping: A systematic review of the literature", *Journal of Data and Information Science*, vol. 2, no. 2, pp. 140, 2017.

[4] E. Chernyak, "An approach to the problem of annotation of research publications." *Proceedings of the 8th ACM international conference on web search and data mining*, ACM, pp. 429-434, 2015.

[5] D. Frolov, B. Mirkin, S. Nascimento, and T. Fenner, "Finding an appropriate generalization for a fuzzy thematic set in taxonomy", Working paper WP7/2018/04, Moscow, Higher School of Economics Publ. House, 60 p., 2018 (URL: *https://wp.hse.ru*/data/2019/01/13/1146987922/WP7_2018_04_____ .pdf, retrieved 17 March, 2019).

[6] Gene Ontology Consortium, "Gene ontology consortium: going forward", *Nucleic Acids Research*, vol. 43, pp. D1049-D1056, 2015.

[7] R. Klavans and K. W. Boyack, "Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?", *Journal of the Association for Information Science and Technology*, 68(4), pp. 984-998, 2017.

[8] D. Lee, R. Cornet, F. Lau, and N. De Keizer, "A survey of SNOMED CT implementations," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 87-96, 2013.

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

[10] B. Mirkin and S. Nascimento, "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.

[11] B. Mirkin, *Clustering: A Data Recovery Approach*, Chapman and Hall/CRC Press, 2012.

[12] G. Mueller and R. Bergmann, "Generalization of Workflows in Process-Oriented Case-Based Reasoning", In FLAIRS Conference, pp. 391-396, 2015.

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

[14] Y. Song, S. Liu, X. Liu, and H. Wang, "Automatic taxonomy construction from keywords via scalable bayesian rose trees," In *IEEE Transactions on Knowledge and Data Engineering*, 27(7), pp. 1861-1874, 2015.

[15] N. Vedula, P.K. Nicholson, D. Ajwani, S. Dutta, A. Sala, and S. Parthasarathy, "Enriching Taxonomies With Functional Domain Knowledge," In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, pp. 745-754, 2018.

[16] J. Waitelonis, C. Exeler, and H. Sack, "Linked data enabled generalized vector space model to improve document retrieval," In *Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC)*, CEUR-WS, vol. 1486, 2015.

[17] C. Wang, X. He, and A. Zhou, "A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances," In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1190-1203, 2017.

 ISBN: 978-1-61208-707-8