Computational Generalization in Taxonomies Applied to: (1) Analyze Tendencies of Research and (2) Extend User Audiences

Dmitry Frolov[1,5], Susana Nascimento[2], Trevor Fenner[3], Zina Taran[4], and Boris Mirkin[1,3]

[1] National Research University Higher School of Economics, Moscow, Russia
[2] Universidade Nova de Lisboa, Caparica, Portugal
[3] Birkbeck, University of London, London, UK
[4] Delta State University, Cleveland, MS, USA
[5] Natimatica, Ltd., Moscow, Russia

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" node in the higher ranks of the taxonomy tree. The head subject is supposed to "tightly" cover the query set, possibly bringing in some errors referred to as "gaps" and "offshoots". Our method, ParGenFS, globally minimizes a penalty function combining the numbers of head subjects and gaps and offshoots, differently weighted. Two applications are considered: (1) analysis of tendencies of research in Data Science; (2) audience extending for programmatic targeted advertising online. The former involves a taxonomy of Data Science derived from the celebrated ACM Computing Classification System 2012. Based on a collection of research papers published by Springer 1998-2017, and applying in-house methods for text analysis and fuzzy clustering, we derive fuzzy clusters of leaf topics in learning, retrieval and clustering. The head subjects of these clusters inform us of some general tendencies of the research. The latter involves publicly available IAB Tech Lab Content Taxonomy. Each of about 25 mln users is assigned with a fuzzy profile within this taxonomy, which is generalized offline using ParGenFS. Our experiments show that these head subjects effectively extend the size of targeted audiences at least twice without loosing quality.

Keywords: Generalization · Fuzzy thematic cluster · Annotated suffix tree · Research tendencies · Targeted advertising.

1 Introduction

The notion of generalization is not absent from the current developments in knowledge engineering and artificial intelligence. Just the opposite. For example, building a supervised classifier fits exactly into the concept of generalization: a classifier generalizes given instances of "yes"-objects into a decision rule to separate the "yes"-class from the rest. This, however, relates to a very special case

at which all the objects are elements of the same variable space. We are going to tackle the case at which we are presented with a crisp or fuzzy subset of different concepts, and one wishes to generalize this subset into a coarser concept tightly embracing the subset. This is, partly, the meaning of the term "generalization" which, according to the Merriam-Webster dictionary, refers to deriving a general conception from particulars. We assume that a most straightforward medium for such a derivation, a taxonomy of the field, is given to us.

Currently, taxonomic constructions mostly concentrate on developing taxonomies, especially those involving referred to in linguistics as hyponymic / hypernymic relations (see, for example, [9, 7]) Also, some activities go in the direction of "operational" generalization: generalized case descriptions involving taxonomic relations between generalized states and their parts are used to achieve a tangible goal such as improving characteristics of text retrieval [8].

This paper does not attempt to develop or change any taxonomy, but rather uses an existing taxonomy. The situation of our concern is a case at which we are to generalize a fuzzy set of taxonomy leaves representing the essence of an empirically observed phenomenon. The rest of the paper is organized accordingly. Section 2 presents a mathematical formalization of the generalization problem as of parsimoniously lifting of a given fuzzy leaf set to nodes in higher ranks of the taxonomy and provides a recursive algorithm leading to a globally optimal solution to the problem. Section 3 describes an application of this approach to deriving tendencies in development of the data science, that are discerned from 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. Its subsections describe 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 lifting results. Section 4 describes an application of the parsimonious generalization method to efficiently extend the audience of targeted advertising over the Internet. More detailed description can be found in [3].

2 Generalization: parsimoniously lifting a fuzzy thematic set in taxonomy

We consider the following problem. Given a rooted taxonomy tree and fuzzy set S of taxonomy leaves, find a node $h(S)$ of higher rank in the taxonomy, that tightly covers the set S.

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 illustrating the situation at which the set of black boxes is lifted to the root. If we accept that set S may be generalized by the root, this would lead to a number, four, of 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 2: here the set is lifted just to the root of the left branch of the tree. The number of gaps here has decreased, to just 1. However, another oddity emerged: a black box on the right, belonging to S but not covered by the node 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. This is less than the number of items emerged at lifting the set to the root (one head subject and four gaps, that is, five), which makes it more preferable. Of course, this conclusion holds only if the relative weight of an offshoot is less than the total relative weight of three gaps.

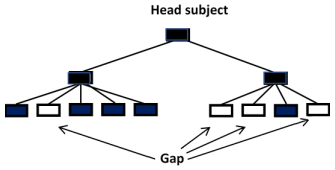

Fig. 1. Generalization of the black box query set by mapping it to the root, with the price of four gaps emerged at the lift.

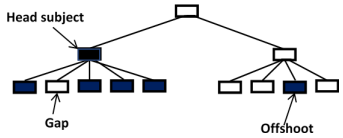

Fig. 2. Generalization of the black box query set by mapping it to the root of the left branch, with the price of one gap and one offshoot emerged at this lift.

We are interested to see whether a fuzzy set S can be generalized by a node h from higher ranks of the taxonomy, so that S can be thought of as falling within the framework covered by the node h. The goal of finding an interpretable pigeonhole for S within the taxonomy can be formalized as that of finding one or more "head subjects" h 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.

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 $u(i) \geq 0$ 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.

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 higher rank 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 (see Figures 1 and 2). 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. For each of these elements a penalty is defined: 1 is the penalty for a head subject, γ, the penalty for a gap, and λ is the penalty for an offshoot.

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 $h \in H - I$.

The penalty function $p(H)$ for a u-cover H is:

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

and we are 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.

First, the tree is pruned from all the non-maximal u-irrelevant nodes. Simultaneously, the sets of gaps $G(t)$ and the internal summary gap importance $V(t) = \sum_{g \in G(t)} v(g)$ in Eq. (1) are computed for each interior node t. 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.

Sets $H(t)$ 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:

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

In case (b), the sets $H(t)$ and $L(t)$ are the unions of those of its children, and $p(t)$ is the sum of their penalties. To obtain a parsimonious lift, whichever case gives the smaller value of $p(t)$ is chosen.

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.

The algorithm ParGenFS is proven to lead to an optimal lifting indeed [3].

3 Highlighting tendencies in research

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. The first pathway tries to discern main categories from the texts, the other, from knowledge of the domain. The first approach is exemplified by clustering and topic modeling; the second approach, by using an expert-driven taxonomy. The main difference between these approaches lies in the level of granularity: the former pathway uses concepts of the same level of granularity as those in texts, whereas the latter approach may bring forth coarser concepts from the higher ranks of a taxonomy.

This paper follows the second pathway by moving, in sequence, through the stages covered in separate subsections 3.1 to 3.6.

3.1 Scholarly text collection

We downloaded a collection of 17685 research papers together with their abstracts published in 17 journals related to Data Science for 20 years from 1998-2017. We take the abstracts to these papers as a representative collection.

3.2 DST Taxonomy

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 the six-layer ACM-CCS 2012 taxonomy of computing subjects [1], which is related to Data Science, and add a few leaves related to more recent Data Science developments The taxonomy itself, with all its 317 leaves, can be found in [3].

3.3 Relevance topic-to text score and co-relevance topic-to-topic similarity index

We first obtain a a keyphrase-to-document matrix R of relevance scores by using the Annotated Suffix Tree approach [2]. This matrix R 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 $a_{ii'}$ between topics i and i' is computed as the inner product of vectors of scores $r_i = (r_{iv})$ and $r_{i'} = (r_{i'v})$. 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 $v = 1, 2, \ldots, V$. A topic is considered relevant to v if its relevance score is greater than 0.2 (a threshold found experimentally, see [2]).

3.4 Fuzzy thematic 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 Fuzzy ADDItive Spectral (FADDIS) clustering algorithm developed in [5] specifically for finding thematic clusters.

After computing the 317×317 topic-to-topic co-relevance matrix, converting in to a topic-to-topic Laplace–transformed similarity matrix [5], and applying FADDIS clustering, we sequentially obtained 6 clusters, of which three clusters are obviously homogeneous. They relate to "Learning", "Retrieval", and "Clustering".

3.5 Lifting the clusters

The three clusters mentioned above are lifted in the DST taxonomy using ParGenFS algorithm with the gap penalty $\lambda = 0.1$ and off-shoot penalty $\gamma = 0.9$.

Lifting Cluster L brings three head subjects: Machine Learning, Machine Learning Theory, and Learning to Rank. Lifting of Cluster R: Retrieval leads to two head subjects: Information Systems and Computer Vision. For Cluster C, 16 (!) head subjects were obtained.

3.6 Drawing conclusions

The "Learning" head subjects show 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 towards learning of ranks and rankings.

Lifting results for the information retrieval cluster R, clearly show: 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, 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". Currently, clustering is not just an auxiliary instrument but rather a model of empirical classification, a big part of the knowledge engineering.

4 Efficient audience extending in targeted advertising

We consider a company, such as start-up Natimatica, Ltd. (see *https://natimatica.com*) associated with us, that supports a service of native advertising. This service follows millions of individual users visiting popular sites providing news, shopping, specific contents, etc. Information of these users is stored in a special system, Data Management Platform (DMP). Each individual user in the DMP is assigned with a subset of the IAB taxonomy of goods and services [4] segments (leaves) relevant to their visits. Each of the segments i is assigned with a real number a_i, a fraction of unity, according to the history of the users visits to the sites under our observation. The totality of the taxonomy segments and membership values assigned to them constitute what can be referred to as the users profile. An advertiser formulates their advert needs as a set of IAB leaves (segments) relevant to the advert. In practice, such a formulation is produced manually by an employee, after a detailed discussion of the advert with the advertiser. A conventional, currently most popular, approach (CAS) requires to pre-specify a threshold t (usually, t=0.3), so that a condition $A(t)$ can be applied: Given a set S of taxonomy segments and a users profile, check if the profile has at least one of the S segments with the value a_i assigned to them so that $a_i > t$. Then the CAS rule requires to expose the advert to all those users for whom condition $A(t)$ holds. An issue with CAS is that the number of users satisfying condition $A(t)$ may be less than the number specified in the advertisement order. In this case, a conventional strategy is to have CAS extended by lessening t to $t^0 < t$, so that more users satisfy condition $A(t^0)$ than $A(t)$ (CASE). In contrast, our strategy, Generalization of User Segments (GUS), is based on the optimal generalizations of user profiles in the IAB taxonomy made off-line. GUS tests condition $A(t)$ by applying it not to the segment concerned but rather to the head subjects.

Table 1 presents comparative results of testing our GUS method at real life advertising campaigns in Natimatica, Ltd. The comparison criteria are: (a) advertising impressions (Imprs in the Table 1) obtained, (b) numbers of clicks, and

(3) click through rates (CTR, (b)/(a)). Our method GUS significantly outperforms the conventional strategies.

Table 1. Advertising campaign results at different targeting methods

Campaign	IAB segments	Metric	CAS	CASE	GUS
Software for parental control of children. Dur. 10 days	Daycare and Pre-School, Internet Safety, Parenting Children Aged 4-11, Parenting Teens, Antivirus Software	Imprs	378933	1017598 (+168.5%)	942104 (+148.6%)
		Clicks	1061	1526 (+43.8%)	2544 (+139.8%)
		CTR,%	0.28	0.15 (-46.4%)	0.27 (-3.6%)
Mortgage at a major Russian bank. Dur. 10 days	Home Financing, Personal Loans	Imprs	159342	275035 (+72.6%)	289308 (+81.6%)
		Clicks	749	853 (+13.9%)	1302 (+73.9%)
		CTR,%	0.47	0.31 (-34.0%)	0.45 (-4.3%)

Acknowledgment. D.F. and B.M. acknowledge continuing support by the Academic Fund Program at the NRU HSE (grant 19-04-019 in 2018-2019) and by the DECAN Lab 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/04516/2019).

References

1. The 2012 ACM Computing Classification System. [Online]. Available: *http://www.acm.org* /about/class/2012 (Accessed 2018, 30 April).
2. E. Chernyak, "An approach to the problem of annotation of research publications", In *Proceedings of the 8th ACM WSDM*, ACM, 429-434, 2015.
3. D. Frolov, B. Mirkin, S. Nascimento, 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, 2018.
4. IAB Tex Lab Content Taxonomy 2017, Available: *https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/* (Accessed 2019, 5 July).
5. B. Mirkin, 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.
6. R. Pampapathi, B. Mirkin, M. Levene, "A suffix tree approach to anti-spam email filtering", *Machine Learning*, 65(1), pp. 309–338, 2006.
7. N. Vedula, P.K. Nicholson, D. Ajwani, S. Dutta, A. Sala, S. Parthasarathy, "Enriching Taxonomies With Functional Domain Knowledge", In *The 41st International ACM SIGIR Conference on R&D in Information Retrieval*, ACM, pp. 745-754, 2018.
8. J. Waitelonis, C. Exeler, H. Sack, "Linked data enabled generalized vector space model to improve document retrieval", In *Proceedings of NLP & DBpedia 2015 workshop and 14th ISW Conference*, CEUR-WS, vol. 1486, 2015.
9. C. Wang, X. He, 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.