A Low Effort Approach to Quantitative Content Analysis

Maria Saburova(✉) and Archil Maysuradze

Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University, 2nd Educational Building, CMC Faculty,
MSU, Leninskie Gory, GSP-1, Moscow 119991, Russia
saburova.mi@yandex.ru, maysuradze@cs.msu.su

Abstract. We propose a workflow for an individual sociologist to be able to use quantitative content analysis in small-scale short-term research projects. The key idea of the approach is to generate a domain-oriented dictionary for researchers with limited resources. The workflow starts like a typical one and then deviates to include content analysis. First, the researcher performs deductive analysis which results in an interview guide. Second, the researcher conducts the small number of interviews to collect a domain-oriented labelled text corpus. Third, a domain-oriented dictionary is generated for the following content analysis. We propose and compare a number of methods to automatically extract a domain-oriented dictionary from a labelled corpus. Some properties of the proposed workflow are empirically studied based on a sociological research on volunteering in Russia.

Keywords: Domain-oriented dictionary · Quantitative content analysis · Term extraction · Low effort sociological workflow

1 Introduction

1.1 Traditional Content Analysis

Content analysis is a research technique for systematic analysis of written communication. Its basic idea is that the large number of words (text units) contained in a piece of text are classified into content categories of interest [6].

Traditional workflow of quantitative content analysis comprises many steps. One of the main concepts of content analysis is a coding scheme — a list of content categories and a procedure which classifies text units into the content categories. Our research focuses on the workflow steps where the coding scheme is developed and utilized.

1. To create a coding scheme, researchers may follow a "deductive" or "inductive" approach. The inductive approach involves automatic derivation of categories based on unlabelled data. The deductive approach is based on comprehensive theoretical considerations. The researchers begin with analytical, or content, categories. These main categories can be further divided

P. Klinov and D. Mouromtsev (Eds.): KESW 2015, CCIS 518, pp. 168–181, 2015.
DOI: 10.1007/978-3-319-24543-0_13

into sub-categories at various levels of hierarchy. It can be said that the coding scheme is the result of the operationalization of the theoretical considerations [8].

In this paper, we are going to address the problem of text unit assignment to categories. That means that sociological projects with predefined categories are easier to adapt to our workflow. The researcher is free to use any approach to develop a list of content categories, and the deductive approach seems to require less effort and less data.

2. Then the researchers define the basic text units to be classified (e.g. individual words, phrases, or paragraphs). Typically researchers tend to classify individual terms in text windows around the objects under study.
3. After that, it is required to develop a code guide. The code guide is essentially a set of categorization rules. Typically the code guide is intended for human coders — experts who perform categorization. A type of the code guide which is convenient for automated processing is a domain-oriented dictionary. The dictionary contains a collection of text units, each unit being exclusively associated with one content category. This dictionary becomes a part of the coding scheme. In the paper we reduce the problem of collecting the dictionary to the problem of text unit assignment to categories based on labelled text units (supervised scenario).
4. When the coding scheme is made, it can be applied to a text corpus by the coders. To apply the coding scheme to the texts in the coherent way the coders follow coding instructions and consult coding examples.
5. The last step of quantitative content analysis is called *quantification*. The occurrences of categorized text units are totalled over the text corpus. There are different software solutions for the step. Automation of the previous steps constitutes a pressing problem.

These steps reveal the complexity of content analysis utilization. It is difficult to develop a coding scheme, especially a dictionary for a code guide. A lot of people are required to analyse large text collections, such as books, essays, news articles, speeches and other written material. For coding step it is also required to teach people to asses text units coherently. In order to secure a good quality of the data a coding workshop should be held, where the coders are familiarized with the coding scheme and made aware of potential pitfalls. And many assessors should be included in this process because of the requirement of quality.

Manual development of the coding scheme often results in a list of questions that should be answered by assessors for each text. Usually the questions are subtle and require the understanding of implicit topics and sentiments, so they cannot be answered automatically using machine learning or natural language processing techniques. On the other hand, this is also an advantage of this method, because it enables researchers to identify implicit concepts and their properties. In contrast, coding using dictionary can be automated. There are many software for quantitative content analysis: Concordance 2.0, Diction 5.0, General Inquirer, TextAnalyst, WordStat v5.0 and so on [14]. Most of the programs use a standard large dictionary (which is inappropriate for domain-specific projects) or demand the dictionary from users. Therefore the main complexity in this type of research is to obtain the domain-oriented dictionary. Our research aims to automate the domain-specific dictionary creation by means of machine learning. To perform it, we had to recognize and solve nonstandard type of machine learning problem — feature distribution among content categories.

1.2 Low Effort Workflow for Individuals

There are many individual researchers who want to use content analysis in their studies. These researchers have little resources and cannot afford many assessors. Their researches are limited in time and conducted in specific field. Therefore individual researchers want to use traditional content analysis, but they have no enough resources or time.

The goal of our work is to provide individual researchers with a low effort content analysis workflow. At present, a typical workflow of an individual researcher includes interviewing of a small number of respondents. We claim that the data collected during the interview design and the interviewing may be used to develop a domain-specific dictionary.

The complete interviewing process includes the following steps [12]:

1. Thematizing: Clarifying the purpose of the interviews and the concepts to be explored.
2. Designing: Laying out the process through which youll accomplish your purpose. This should also include ethical considerations.
3. Interviewing: Doing the actual interviews.
4. Transcribing: Creating a written text of the interviews.
5. Analyzing: Determining the meaning of the information gathered in the interviews in relation to the purpose of the study.
6. Verifying: Examining the reliability and validity of the information gathered.
7. Reporting: Telling others what you have learned or discovered.

Let us compare the two workflows. On one hand, there is a labour-intensive content analysis method and computer programs developed specifically for this process. However, such programs require a domain-oriented dictionary. On the other hand, there are low effort sociological researches that collect interviews. We are going to use the collected interview data as a labelled text corpus.

These workflows have steps analogous with each other. Interview design corresponds to the deductive step of the coding scheme creation in content analysis. Another similarity between the semi-structured interview analysis and the content analysis is that main categories and particular questions are selected during coding scheme development. A question list is composed as a result of coding scheme analysis and can be used for the interviewing process. However, in small sociological researches it is common to summarize materials obtained from a small number of respondents only. If the researches were given a tool to process large data timely, they would study large text corpora.

To easily use content analysis algorithms in low-effort researches, we propose a workflow when an individual researcher proceeds to automatic quantitative content analysis after interviewing and transcribing steps. To make it possible, we propose a method of automatic construction of a domain-specific dictionary that only uses data collected during the interviewing. The method can be performed on a regular personal computer.

The rest of the paper is organized as follows. In Section 2 we formalize the notion of dictionary. In Section 3 we review related work. Section 4 describes particular qualities of the problem. Section 5 introduces a tripartite data model which underlies our formal constructions. The mathematical description of proposed methods resides in Section 6. In Section 7 we describe experimental setup and discuss the results. Section 8 concludes the paper.

2 Dictionary: The Lexical Core of Categories

In our research, a dictionary is a list of the units that should be unambiguously assigned to one of the categories. A part of dictionary, which related to one category, we perceive as lexical core of category. This idea is formalized by the following statements:

1. The presence of units is the text marker. Then marker is a binary feature.
2. Dictionary unit presence is a category reference indicator. Some markers are assigned to a class and each of them are assigned directly to one class.
3. We consider the problem of creating a dictionary as the problem of distributing the features, which occurs in data mining.

The problem of feature distribution by classes consists in the follows. Given a training set in which each object has a feature description and a class label from a finite set of predefined classes. Requires every feature assigned uniquely to either one of the predefined classes or special additional class. Interpretation of additional class depends on the subject area.

Note that in this formulation there is no initial marking of features relating to classes. Such markings can be used for quality evaluation of the solution.

In the texts categorization area [17] and the content analysis [7,18] the considered problem can be interpreted as the problem of lexical (semantic) core of category separation. In this case, the objects are text fragments, classes are text collections categories, features show the presence of lexical markers in texts (e.g., single words, phrases, specific terms). Such descriptions are typical for the bag-of-words and vector space models [13]. The additional class are sometimes referred to a common vocabulary class.

3 Related Work

3.1 Lexical Core Definition

Lexical core of a category is defined as a set of such lexical markers (words, phrases, terms) that their presence in the text clearly assigns the text to this category. In that way, the markers from the lexical core of category are significant for this category and at the same time are not significant for the text assignment to other categories. In this work we do not consider the case where a lexical marker is significant for multiple categories. As well, "negative" markers are not considered too. Negative marker is a marker such that its presence or absence prohibits to assign the text to a certain category. If the word presence forbids a class, these negative markers are called negative keywords. Excluding negative markers from the study are due to the fact that they are not well-established in data mining cases because of reducing the generalization ability.

To better describe the concept of lexical core, it is useful to compare it with the semantic core, which is used in the subject area of search engine optimisation. Semantic core of object is the set of markers that describes the object and which reference in the search request leads to object selection in the search results. For example, site position in search engine results depends on the completeness and accuracy of the semantic core development [3]. Notice the following differences between lexical and semantic kernels. First, in our work we talk about class cores while in the field of search engine optimisation focuses on the object cores. Secondly, we work with markers that are determined by the objects creators (e.g. author keywords), and in the field of search engine optimisation we work with markers used by potential object users. The differences between the lexical and semantic core are concerned with input data, but the prospective outcomes (i.e. markers sets) are similar. Therefore, it is rational to consider both formulations.

3.2 Lexical Core vs. Semantic Core

In the literature the term "lexical core" is used in various other senses that are not relevant to this work. To avoid displacement of concepts, we list these alternative meanings. *Lexical language core* is often mentioned in linguistics and is defined as the set of core lexical units that are opposed to the peripheral lexical units [11]. *Semantic kernel* is used in machine learning as a *meaningful* measure of closeness in the documents space [4][9]. In the information retrieval area, the term *semantic core* is defined as a special data structure, that describes the contents of the document. This semantic core is a semantic network with vertices (keywords and phrases selected from the document), and the weight between connection nodes is calculated based on some similarity measure [5].

3.3 Quality Measures for Lexical Cores

In the above mentioned papers on lexical/semantical kernels the quality scores are usually specified for the main problem and not for the lexical kernel itself.

Involving assessors is another method to control the quality of the dictionaries. Assessors evaluate words that were selected to the dictionaries. There are two kinds of this evaluation process.

In first method assessors receive a list of all units and a list of all categories. They should choose one category for every unit. The second method is devoted to the quality assessment of the words marked by the algorithm.

3.4 Learning Topic Models

Topic modeling area is quite close to out research. Although basic problem statements have important differences: topic modeling is concerned with clustering of words and documents by topics that are unknown in advance. The topic model of text documents collection determines how topics are distributed in each document and which words (terms) determine every topic [20]. An important difference between our problem statements and topic modeling statements is that the number of topics (and especially their list) is not known in advance in most applications and is one of the most important parameters for setting up a topic model. Also, unlike the studied problem, in the topic modeling statement it is required to find the distribution of each word on all topics. That is, the probabilities of belonging to each cluster are determined for every feature. However, there are methods, that add the requirement of unambiguity the word-topic relation, for example, methods of regularization. The combination of regularizers usually requires structural adjustment of model [10], although more efficient approaches were proposed recently [19].

There are topic quality measures that are used in topic modeling and can be used to evaluate dictionary quality. Such measures are usually internal or external evaluations of topic coherence (interpretability) [16]. Topic interpretability is estimated subjectively. Topic quality measure should be well correlated with *interpretability by experts* measure. The pointwise mutual information method is recommended in [15]. The general idea is that the topic is called coherent if the most frequent topic words appear close to each other nonrandomly in the collection. In the current research we can consider dictionary coherence. If the words in the dictionary are ranked, then we can consider the coherence of the dictionary top part (for example, top-10, top-100). Also a part of words assigned to the common vocabulary can be considered as a quality measure.

Anchor words idea was suggested [2] in topic modeling recently. Anchor word has non-zero probability only in one topic. If a document contains this anchor word, then it is guaranteed that the corresponding topic is among the set of topics used to generate the document [1]. Therefore the set of anchor words can be interpreted as topic dictionary.

4 Motivation

Let us describe the properties of considered problem.

1. In our problem statements the results will obtain their own interpretation and value. In similar research considered in literature features are assigned to classes only to increase classification quality.
2. We are required to make a decision on each feature, i.e. it is necessary to distribute all features among categories. This condition is not required in most of publications. Usually most of features are rejected.

3. In our research each feature should be labelled uniquely. As opposed, in all publications each feature can be labelled with a few classes.
4. The ratio of the number of features to the number of precedents is much greater than in traditional classification problems.
5. We interpret each feature as a positive assessment of the corresponding properties existence. In particular, the marker is interpreted as the presence of some object properties.
6. In our problem statement features have a Boolean type or can be naturally reduced to this type.

Consequently, the problem of feature distribution among classes in the described form has not been investigated previously in the literature. Addressing this gap and proposing methods for solving the problem is the goal of this paper.

Note that our assignment of every feature to exactly one class has no semantic basis; we perform such assigment based on purely statistical properties of the feature. This approach can result in markers that can fit several categories according to common sense. However, our experiments show that most of markers are very reasonable and accepted by experts.

5 Tripartite Data Model

In this study we focus on the problems where features are markers or measures of some properties. Data in such problems can be naturally represented by the relational model using a set of binary relations. We call such model *tripartite* and specified its three components: objects, markers and classes. We also define three binary heterogeneous relations between the units of analysis. The relation is called heterogeneous if it connects different units of analysis.

This model is convenient for formalizing and solving many applied analytical tasks, in particular in sociology, scientometrics, text processing and other fields. Different problems can be formalized according to this model: object classification, class definition of the new object and many others, including markers distribution to the classes. Many problems are defined according to the scheme where two relations are given, and the third should be restored.

The tripartite model formalization can be clarified for the problem of marker distribution among the classes. In this case the relation between markers ans classes becomes functional – it partially maps markers to the original classes. A model where some binary relations are partial mappings, will be called the *tripartite semihard model.*

In the problem of assignment classes to features the binary relation between objects and classes is a (total) mapping, and the relation between objects and markers is arbitrary (many to many). Degenerate situation where the object is not associated with any marker or the marker is not associated with any object, will not be considered. In other words, the relation between documents and markers should be left-full and right-full.

6 Dictionary Construction Methods

6.1 Function as a Classifier Argument

In the studied problem statement, the markup is given for objects rather than features. Therefore, we cannot immediately state the problem of markers classification and should start with reducing the problem of feature classification to the task of object classification. For this reason we propose an information model where partial function from attributes to the classes is included explicitly as a parameter. In other words, the classifier assigns the feature to the class. Required partial function will be obtained automatically after fitting classifier of this model to the data.

We introduce the following notation:

1. T — the number of features, t — feature number from 1 till T,
2. I — the number of labeled objects, i — object number from 1 till I,
3. J — the number of classes, j — class number from 1 till J,
4. a_t — class number, which is mapped with feature t, required function,
5. f_{it} — value of feature t for object i, in particular, 0 or 1 for binary relation object-feature,
6. c_i — real object label i.

We use linear information model as one of the simplest. We introduce non-negative feature's weight, which is marked as w_t. It is required to classify the object with features $f_1,\dots, f_T$. The estimate of object belonging to the class k is calculated as the sum of the weights products and values of those features t, which is assigned to this class:

$$\Gamma_k = \sum_{t=1}^{T} w_t f_t [a_t = k].$$

We interpret the feature values here as positive measures of some properties. Decision rule assigns the object to the class with the highest rating:

$$A = argmax_k \Gamma_k.$$

6.2 Multiclass SVM Analogue Method

We describe now learning method for algorithm from the information model. Parameters $\{a_t\}$ and $\{w_t\}$ will be configured by solving margin maximizing problem. Introduced method has lots of similarities with multiclass SVM. Margin is defined as the difference between the score for the real class and the maximum estimation among the other classes.

$$\frac{1}{2}||w||^2 + C\sum_{i,j} \xi_{ij} \rightarrow \min_{w,\{\xi_{ij}\}_{i,j}} \tag{1}$$

$$\sum_t w_t f_{it}([a_t = c_i] - [a_t = j]) \geq 1 - \xi_{ij},\ \forall i,\ \forall j \neq c_i \tag{2}$$

$$w_t \geq 0,\ \forall t \tag{3}$$

$$\xi_{ij} \geq 0,\ \forall i,\ \forall j \neq c_i \tag{4}$$

Slack variables ξ_{ij} are introduced here to deal with the case when classes are not linearly separable. When sample is linearly separable, problem can be simplified:

$$\begin{array}{l} \frac{1}{2}||w||^2 \to \min \\ \sum_t w_t f_{it}([a_t = c_i] - [a_t = j]) \geq 1,\ \forall i,\ \forall j \neq c_i \\ w_t \geq 0,\ \forall t \end{array} \tag{5}$$

Dual problem is defined for solving this problem.

$$\begin{array}{l} \sum_{i,j} \alpha_{ij} - \frac{1}{2}||\beta_t + X_{ijt}^T \alpha_{ij}||^2 \to \max \\ 0 \leq \alpha_{i,j} \leq C, \\ \beta_t \geq 0,\ \forall t \end{array} \tag{6}$$

Here α_{ij} и β_t is dual variables, X_{ijt} is defined as

$$X_{ijt} = f_{it}([a_t = c_i] - [a_t = j]). \tag{7}$$

After dual problem solving, it is needed to return to initial features.

$$\begin{array}{l} \beta_t = 0, where X_{ijt}^T \alpha_{ij} \geq 0 \\ \beta_t = -X_{ijt}^T \alpha_{ij}, where X_{ijt}^T \alpha_{ij} < 0. \end{array} \tag{8}$$

Initial features can be find from formula:

$$w_t = \beta_t + X_{ijt}^T \alpha_{ij}. \tag{9}$$

The dual problem is a linear programming problem, when a_i are fixed. Interior-point method can be used to solution this problem.

Coordinate descent method is used to train a_i:

1. The algorithm starts from initial point: every feature is assigned to the class in which it often occurs.
2. On each iteration random feature s is selected. For this feature look over all classes for which this feature vote.
3. Weights w_t are optimized for each of these classes, when a_t are held.
4. Class a_s with the maximal value of the dual problem functional is assigned to the feature.
5. The procedure is repeated until convergence or until the specified number of iterations will be reached.

6.3 One-vs.-one SVM in Relation to the Multiclass Problem

Here we propose another dictionary development algorithm. This algorithm based on the same idea of feature distribution during object classification. We consider SVM algorithm for binary classification problem:

$$\begin{array}{l} \|w\|^2 + C\sum_{i=1}^{\ell} \xi_i \to \min_{w,\xi}; \\ y_i\langle w, x_i\rangle \geq 1 - \xi_i, \quad \forall i = 1, \ldots, \ell; \\ \xi_i \geq 0, \quad \forall i = 1, \ldots, \ell. \end{array} \tag{10}$$

The problem of multiclass classification can be reduced to the set of binary classification problems. We use one-vs.-one scheme:

1. We train binary classifiers a_{sk} for all classes pairs $s \neq k$;
2. Each of them distinguishes documents of class s from documents of class k;
3. Weights w_t^{sk} is considering to each classifier;
4. If $w_t^{sk} > 0$, then feature t vote for class s, else k;
5. Feature t is assigned to class s, if $w_t^{sk} > 0$ for more than a half pairs $k \neq s$.

7 Experiments

7.1 Data Description

We chose responses to interview questions for our experiments. The data was obtained from the project “Resource of avantgarde groups volunteerism for Russian modernization”, which was implemented by the Fund “Public opinion” in collaboration with researchers with the use of state support funds allocated by the Institute for public planning grant in accordance with the decree of the President of the Russian Federation from 02 March 2011 No. 127rp[1].

The data consists of 20 interviews with leaders of volunteering organizations. Interview categories are: “Supervisor portrait”, “Objectives and content of the organization’s activities”, “The Concept of volunteerism”, “Working with volunteers”, “Volunteers portrait”, “Incentives and barriers to volunteering activities”.

Each document was divided into 6 sections in accordance with the categories. Each section was considered as one object, for a total of 120 objects. Each interview category represents one class. We have normalized each word in the text (note that different Russian words can have similar English translation, which is the result of language specifics and not because of normalization problems). Stop-words were not excluded for two reasons: (1) methods were designed in such way that stop-words should be assigned to special additional class and (2) presence of stop-words in our domain-specific texts may have some signal. After text normalization and duplicates filtering the corpus consists of 7241 word.

Assessor labeling was used for dictionary quality evaluation. Experts reviewed each word in the dictionary and evaluated a relation between the word and

[1] The data can be provided by the Fund “Public Opinion” on request: fom@fom.ru

Table 1. Dictionary produced by "Multiclass SVM analogue" method. There are 6 categories. For each category the first 18 words are shown. Words are ordered by their score (not shown). For each word experts assessed whether the word belongs to the category. The columns display Russian transliteration, English translation, and precision at corresponding level. Note that different Russian words may have equal English translations. Russian words are normalized.

Supervisor portrait			Objectives and content of the organization's activities			The Concept of volunteerism		
special'nost'	specialty	100%	finansirovanie	financing	100%	nazyvaju	call	100%
uchus'	learn	100%	itog	summary	100%	nazvat'	call	100%
skol'ko	how much	67%	reshenie	solution	100%	razovyj	one	100%
nemnogo	a little	50%	budushhee	future	100%	ponjatie	the concept	100%
sozdanie	creation	60%	istochnik	source	100%	obshhestvennik	public man	100%
gde	where	67%	voznikaju	arise	83%	bezvozmezdnyj	free	100%
ozhidanie	waiting	71%	zasluga	merit	86%	aktivist	activist	100%
universitet	University	75%	poslednij	last	75%	inogda	sometimes	100%
okonchanie	the end	78%	naibolee	the most	67%	znachimyj	significant	100%
god	year	70%	reshit'	to solve	70%	sistematicheskij	systematic	100%
reshil	decided	73%	cel'	goal	73%	dobrovol'chestvo	volunteering	100%
davno	long	75%	trudnost'	the difficulty	75%	social'no	social	100%
lichno	personally	77%	vlast'	power	77%	jepizodicheskij	episodic	100%
opravdalsja	justified	79%	vtoroj	second	71%	darit'	to give	100%
objazannost'	duty	80%	sposob	method	73%	schitaju	think	100%
rasskazal	told	81%	postavit'	to put	75%	besplatnyj	free	100%
institut	Institute	82%	stavlju	put	71%	mezhdu	between	94%
posle	after	78%	reshat'	to solve	72%	opredelenie	definition	94%
Working with volunteers			Volunteers portrait			Incentives and barriers to volunteering activities		
shtatnyj	staffing	100%	chashhe	more often	0%	meshaju	disturb	100%
dovolen	happy	100%	muzhchina	man	50%	otnoshus'	am	50%
navyk	skill	100%	zhenshhina	woman	67%	municipal'nyj	municipal	67%
special'nyj	special	100%	stanovljus'	become	75%	gosudarstvennyj	state	75%
proishozhu	happen	80%	molodoj	young	80%	prestizhen	prestigious	80%
internet	Internet	83%	dumaju	think	67%	naselenie	population	83%
pishu	write	86%	dobryj	good	71%	bol'shinstvo	most	86%
vazhno	important	88%	starshe	older	75%	modno	fashionable	88%
obojtis'	do	89%	sluchaj	case	67%	struktura	structure	89%
obraz	the way	90%	edinyj	single	60%	strana	country	90%
lichnyj	personal	91%	portret	portrait	64%	kazhetsja	it seems	82%
privlekat'	to attract	92%	procent	percentage	67%	doverie	trust	83%
meroprijatie	the event	92%	narisovat'	draw	69%	gorod	the city	85%
dorog	roads	86%	zhena	wife	64%	resurs	resource	86%
jetap	stage	87%	duh	the spirit	60%	ispol'zujushhij	using	87%
bez	without	81%	politicheskij	political	63%	dobrozhelatelen	friendly	88%
professional'nyj	professional	82%	religioznyj	religious	65%	doverjaju	trust	88%
sovmestnyj	joint	83%	blagopoluchnyj	safe	67%	biznes	business	89%

obtained category as 0 or 1. The main difficulty here is that it is hard to evaluate a category for the word without any context. We say that the word belongs to a category if there is a context in which the word is consistent with category name, questions, and comments of this category from interview guide.

7.2 Experimental Results and Discussion

Experimental results represented in 1 are obtained from the first method, which was called "Multiclass SVM analogue" method. Experiment results represented in 2 are obtained from the second method — One-v.s.-one SVM. In these tables the first 18 words of obtained dictionaries are represented. Experts reviewed each word in the dictionary and evaluated a relation between the word and obtained category as 0 or 1; the votes were aggregated by majority vote. Each method allows to calculate an importance of the word:

1. The "Multiclass SVM analogue" method infers a weight for each feature, and we take this weights as feature importances.
2. The "One-v.s.-one SVM" method infers feature weights for each pair (s, k) of classes. We average this weights over all class pairs with $w_t^{sk} > 0$ for each feature, and use this average as an importance value.

The exact value of feature importance has no significant interpretation; we use it only to define the order of words inside each class. Based on this evaluation we

Table 2. Dictionary produced by "One-v.s.-one SVM" method. There are 6 categories. For each category the first 18 words are shown. Words are ordered by their score (not shown). For each word experts assessed whether the word belongs to the category. The columns display Russian transliteration, English translation, and precision at corresponding level. Note that different Russian words may have equal English translations. Russian words are normalized.

Supervisor portrait			Objectives and content of the organization's activities			The Concept of volunteerism		
special'nost'	specialty	100%	itog	summary	100%	aktivist	activist	100%
uchit'sja	to learn	100%	reshenie	solution	100%	nazyvat'	call	100%
skol'ko	how much	100%	finansirovanie	financing	100%	razovyj	one	100%
gde	where	100%	istochnik	source	100%	obshhestvennik	public man	100%
rasskazat'	to tell	100%	budushhee	future	100%	inogda	sometimes	100%
nemnogo	a little	83%	naibolee	the most	83%	social'no	social	100%
posle	after	71%	cel'	goal	86%	znachimyj	significant	100%
sozdanie	creation	75%	trudnost'	the difficulty	88%	ponjatie	the concept	100%
ozhidanie	waiting	78%	zasluga	merit	89%	vozmezdnyj	reimbursable	100%
zanjat'sja	to do	80%	postavit'	to put	90%	jepizodicheskij	episodic	100%
okonchanie	the end	82%	poslednij	last	82%	schitat'	take	100%
universitet	University	83%	sposob	method	83%	sistematicheskij	systematic	100%
potom	then	77%	zametnyj	noticeable	77%	dobrovol'chestvo	volunteering	100%
opravdat'sja	excuses	79%	vlast'	power	79%	platnyj	paid	93%
lichno	personally	80%	voznikat'	to occur	73%	ljuboj	any	87%
god	year	75%	reshat'	to solve	75%	dobrovolec	volunteer	88%
nachat'	to start	76%	novyj	new	71%	mezhdu	between	82%
zakonchit'	finish	78%	vtoroj	second	67%	volontjor	volunteer	83%

Working with volunteers			Volunteers portrait			Incentives and barriers to volunteering activities		
shtatnyj	staffing	100%	chastyj	frequent	0%	meshat'	disturb	100%
navyk	skill	100%	muzhchina	man	50%	otnosit'sja	apply	50%
dovol'nyj	happy	100%	zhenshhina	woman	67%	dobrozhelatel'nyj	friendly	67%
special'nyj	special	100%	portret	portrait	75%	bol'shinstvo	most	75%
internet	Internet	100%	molodoj	young	80%	municipal'nyj	municipal	80%
obojtis'	do	100%	duh	the spirit	67%	gosudarstvennyj	state	83%
jetap	stage	100%	edinyj	single	57%	prestizhnyj	prestigious	86%
kontrolirovat'	control	100%	stanovit'sja	to become	63%	naselenie	population	88%
zatrata	cost	100%	blagopoluchnyj	safe	67%	doverie	trust	89%
stimulirovat'	to stimulate	100%	aktivnyj	active	70%	modno	fashionable	90%
privlekat'	to attract	100%	cennost'	value	73%	razvitie	development	91%
dorogoj	dear	92%	starshij	senior	75%	strana	country	92%
bez	without	85%	jekonomicheski	economically	77%	struktura	structure	92%
privlech'	to attract	86%	apolitichnyj	apolitical	79%	kollega	colleague	93%
vazhno	important	87%	princip	the principle	73%	biznes	business	93%
sotrudnik	employee	88%	gruppa	group	75%	doverjat'	trust	94%
pisat'	write	82%	iskat'	search	71%	kazhetsja	it seems	88%
proishodit'	to happen	78%	vozrast	age	72%	soobshhestvo	community	89%

count the precision at each level. Recall value could also be of interest, but its estimation is too expensive in our setting (the full ground truth domain-oriented dictionary should be build by experts). Dictionaries from two methods are quite similar but not the same. Both of them are consistent with expert's opinion and can be used as a domain-oriented dictionary in low effort sociological workflow.

The first method is very memory-intensive and requires to solve a non-trivial quadratic programming problems which can be time consuming. From that perspective the second method appears to be more suitable in practice.

8 Conclusion

We proposed and implemented a low effort sociological workflow that allows individual researchers to use the quantitative content analysis. The main challenge in this type of research is to obtain the domain-specific dictionary. At present, it is common for an individual researcher to interview respondents. Our technique makes it possible to collect the dictionary from interview data. So, after the interviewing the individual researcher can proceed to qualitative content analysis. The technique may be run on a regular personal computer. The problem of dictionary construction is formalized in terms of feature distribution and two original solutions are proposed. Proposed methods were implemented and tested on real data. Experiment results were consistent with the expert opinion. Future work will shift the focus to user interaction.

Acknowledgments. This work is partially supported by the Russian Foundation for Basic Research projects No. 13-01-00751a and No. 15-07-09214a.

References

1. Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. arXiv preprint arXiv:1212.4777 (2012)
2. Arora, S., Ge, R., Moitra, A.: Learning topic models-going beyond svd. In: 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS), pp. 1–10. IEEE (2012)
3. Arsirij, E., Antoshhuk, S., Ignatenko, O., Trofimov, B.: Avtomatizacija razrabotki i obnovlenija semanticheskogo jadra sajta s dinamicheskim kontentom. Shtuchnijintelekt (2012)
4. Basili, R., Cammisa, M., Moschitti, A.: A Semantic Kernel to Classify Texts with Very Few Training Examples. Informatica (Slovenia) **30**, 163–172 (2006)
5. Baziz, M., Boughanem, M., Aussenac-Gilles, N.: Conceptual indexing based on document content representation. In: Crestani, F., Ruthven, I. (eds.) CoLIS 2005. LNCS, vol. 3507, pp. 171–186. Springer, Heidelberg (2005)
6. Bengston, D.N., Xu, Z.: Changing national forest values: a content analysis. Research Paper NC-323. St. Paul, MN: US Dept. of Agriculture, Forest Service, North Central Forest Experiment Station (2006)
7. Berelson, B.: Content analysis in communication research (1952)

8. von dem Berge, B., Poguntke, T., Obert, P., Tipei, D.: Measuring intra-party democracy
9. Cristianini, N., Shawe-Taylor, J., Lodhi, H.: Latent semantic kernels. Journal of Intelligent Information Systems **18**(2–3), 127–152 (2002)
10. Khalifa, O., Corne, D.W., Chantler, M., Halley, F.: Multi-objective topic modeling. In: Purshouse, R.C., Fleming, P.J., Fonseca, C.M., Greco, S., Shaw, J. (eds.) EMO 2013. LNCS, vol. 7811, pp. 51–65. Springer, Heidelberg (2013)
11. Kuznecov, A.M.: Strukturno-semanticheskie parametry v leksike: na materiale anglijskogo jazyka. Nauka (1980)
12. Kvale, S., Brinkmann, S.: Interviews: Learning the craft of qualitative research interviewing. Sage (2009)
13. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
14. Neuendorf, K.: Computer content analysis programs (2015). http://academic.csuohio.edu/kneuendorf/content/cpuca/ccap.html (Accessed July 13, 2015])
15. Newman, D., Karimi, S., Cavedon, L.: External evaluation of topic models. In: Australasian Doc. Comp. Symp., 2009. Citeseer (2009)
16. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Association for Computational Linguistics (2010)
17. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34**(1), 1–47 (2002)
18. Stemler, S.: An overview of content analysis. Practical Assessment, Research & Evaluation **7**(17), 137–146 (2001)
19. Voroncov, K.V., Potapenko, A.A.: Reguljarizacija verojatnostnyh tematicheskih modelej dlja povyshenija interpretiruemosti i opredelenija chisla tem. Mezhdunarodnaja konferencija po komp'juternoj lingvistike "Dialog", pp. 676–687 (2014)
20. Vorontsov, K., Potapenko, A.: Pregularization, robustness and sparsity of probabilistic topic models. Computer Research and Modeling **4**(4), 693–706 (2012)