Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 193 – 200

Procedia Computer Science

Information Technology and Quantitative Management , ITQM 2014

A Method for Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources

Ekaterina Chernyak[a,*], Boris Mirkin[b]

National Research University Higher School of Economics, Moscow, Russia

[a] *echernyak@hse.ru*
[b] *bmirkin@hse.ru*

Abstract

A two-step approach to taxonomy construction is presented. On the first step the frame of taxonomy is built manually according to some representative educational materials. On the second step, the frame is refined using the Wikipedia category tree and articles. Since the structure of Wikipedia is rather noisy, a procedure to clear the Wikipedia category tree is suggested. A string-to-text relevance score, based on annotated suffix trees, is used several times to 1) clear the Wikipedia data from noise; 2) to assign Wikipedia categories to taxonomy topics; 3) to choose whether the category should be assigned to the taxonomy topic or stay on intermediate levels. The resulting taxonomy consists of three parts: the manully set upper levels, the adopted Wikipedia category tree and the Wikipedia articles as leaves.Also, a set of so-called descriptors is assigned to every leaf; these are phrases explaining aspects of the leaf topic. The method is illustrated by its application to two domains: a) Probability theory and mathematical statistics, b) "Numerical analysis" (both in Russian).

Keywords: taxonomy refinement; string-to-text relevance; utilizing Wikipedia; suffix trees.

1. Introduction

Taxonomy, or hierarchical ontology, is a popular computational instrument for representation, maintaining and usage of domain knowledge [10,13]. A taxonomy is a rooted tree formalizing a hierarchy of subjects in an applied domain. Such a tree corresponds to a generalizing relation between the subjects such as B is part of A or A is more general than B. Automatization of taxonomies building is important for further progress of computational text processing as well as for improving information retrieval [14,17]. The mainstream work for advancing into the problem assumes usage of a large collection of unstructured texts related to the domain. These are used to find a set of keywords/keyphrases with a clear cut relation of inheritance between them so that the set of keywords and the relation are output as the taxonomy that has been looked for. Drawbacks of this approach are well-known: (a) not every domain can be supplied with a

* Corresponding author. Tel.: +7(495) 772-95-90**22668
E-mail address: echernyak@hse.ru

doi:10.1016/j.procs.2014.05.260

representative large corpus of unstructured text documents, and (b) methods for finding semantic relations between words are not that perfect currently so that both the vocabulary and structure of a found taxonomy are less than satisfactory, as a rule[9]. Therefore, the idea of using an Internet resource, such as Wikipedia, instead seems quite natural[12]. Moreover, one should expect that Wikipedia would supply the taxonomist with a set of subjects and a hierarchic relation over them because of its very nature. Yet one cannot expect that the subjects and the hierarchy can be transferred for the task as easy as it seems to be. The issue is that Wikipedia writers are more enthusiastic than professional. Therefore, one should expect that either the set of subjects or the hierarchy or even some articles or all of those no one can say what may be flawed.

In the remainder, we describe a semi-automatic method for deriving a domain taxonomy in two steps. First step, manually building a frame, top level, taxonomy, usually of one or two layers only, by taking them from the official documents and definitions. Second step is of step-by-step refining the taxonomy topics by adding fragments of the Russian Wikipedia category tree, and articles in the categories, both pre-filtered of noise items. A string-to-text relevance score, based on annotated suffix trees, is used throughout. The method of refining of a taxonomy leaf includes the following stages, after the relevant materials from Wikipedia are downloaded: for every taxonomy topic we find Wikipedia relevant categories and articles and refine the topic by found Wikipedia entities. The method is illustrated by its application to two mathematics domain, Probability theory and mathematical statistics and "Numerical analysis" (in Russian), which shows both advantages and drawbacks of the method.

This application is relevant to our work on using taxonomies for computational visualization and interpretation of texts paper abstracts and course syllabuses in the field of applied mathematics and informatics. In Russian, the only publicly available taxonomy of Mathematics and related areas is the classification for the government-sponsored Abstracting Journal of Mathematics[15] developed in 1999. This is somewhat outdated and unbalanced. Fortunately, in Russia one can find a live and frequently updated classification of sciences maintained by the High Attestation Committee (HAC) of Russia supervising the national system of PhD and ScD theses[7]. It is not quite deep covering just two layers of the body of science. Two or three more layers can be derived from the so-called specialty passports available for each of the classification leaves. Yet all these layers are of rather coarse granularity. To reach the base granularity concepts, such as the concept of derivative in mathematics, one needs two to four layers of more and more refined concepts.

This specifies the problem. We need a method to refine a coarse taxonomy by using Wikipedia (ru.wikipedia.org). The method should allow as to produce a more or less balanced tree structure. One more requirement to the refinement method is that every refined leaf in its output is to be assigned with a number of keywords or key phrases clarifying the contents of the corresponding concept. Such is the ACM Computing Classification System[1] so that we refer to the required balance properties and clarifying labels as the ACM CCS gold standard.

The problem of refinement of a taxonomy has received some attention in the literature. A big question arising before any refinement starts is about the sources for generating new topics. Usually the results of a search engine query, such as A consists of..., where A is an existing taxonomy topic, are analyzed[16]. Such a query would lead to a set of concepts that can be considered as potential subtopics for topic A. This works especially easy if the ontology is represented by means of a formal language, such as OWL, by introducing new logical relations[5]. On the whole, not only fully unstructured sources like collections or corpora of text may work well in this situation, but also sources such as other taxonomies or ontologies can be used. Another approach, becoming much popular, is using the Wikipedia as a major source of new topics[4,12,16,18]. Wikipedia offers a lot of data types, such as unstructured texts, images, the category trees, revision history, redirect pages and covers many specific knowledge domains.[12] lists some advantages of Wikipedia usage for any kind of taxonomy construction:

- Wikipedia is consistently updated, thus Wikipedia-based taxonomies might be easily maintained.
- Wikipedia is multilingual, so any method developed for one languages can be transferred to another.

In papers[4,12,16,18]different approaches for constructing[4,12] or refining[16,18] ontologies and taxonomies by using Wikipedia article data are presented. In Ponzetto the Wikipedia articles are used as a source of topics, in[18] the Wikipedia category tree, in[4] both the articles and the category labels, and in[16] the Wikipedia infoboxes are utilized. Our approach to refining taxonomies is somewhat different. We extract topics both from the Wikipedia category tree and from the articles. This allows us to follow the ACM-CCS gold standard of taxonomy. By restricting the domain

of the taxonomy to smaller topics such as the probability theory and mathematical statistics, we avoid the issue of big Wikipedia data and, also, get the possibility to manually examine the results.

The reminder is organised as follows: section 1 provides details of our approach, section 2 presents main results and some issues to be tackled in the future. The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2014.

2. Our approach

First we specify the taxonomy domain and manually form the frame of the taxonomy by extracting basic topics from the publicly available instruction materials of the Higher Attestation Commission (HAC) of Russia. The data for refining the taxonomy frame is extracted from Wikipedia. We will provide two examples of refined taxonomies: of 1) probability theory and mathematical statistics and 2) numerical analysis. The frames of both taxonomies are three-levels rooted trees of the main topics in the domain (see Tables 1 and 2, correspondingly).

Table 1: Probability theory and mathematical statistics taxonomy frame

1	Probability theory
1.01	Models and characteristics of random events
1.02	Probability distributions and limit theorems
1.03	Combinatory and geometrical probability problems
1.04	Random processes and fields
1.05	Optimization and algorithmic probability problems
2	Mathematical statistics
2.01	Methods of statistical analysis and inference
2.02	Statistical estimators and estimating parameters
2.03	Test statistics and statistical hypothesis testing
2.04	Time series and random processes
2.05	Machine learning
2.06	Multivariate statistics and data analysis

Table 2: Numerical analysis taxonomy frame

1	Numerical analysis
1.01	Algorithms for numerical problem solving
1.02	Numerical method for applied problems
1.03	Software for numerical methods
1.04	Numerical analysis theory
1.04.01	Properties of algorithms
1.04.02	Algorithmic efficiency
1.04.03	Validation of algortithms

The next step is to define corresponding Wikipedia categories. For each domain we choose only category of the same name, so there is no need to address any other categories. Among variety of Wikipedia content we will exploit inly two data types:

- The hierarchical structure of Wikipedia category tree
- The collection of unstructured Wikipedia articles.

Hereafter we are going to use the Wikipedia category tree for extending our taxonomy tree. We try to assign some Wikipedia categories to every taxonomy topic of the first and second levels. First, we find those Wikipedia categories

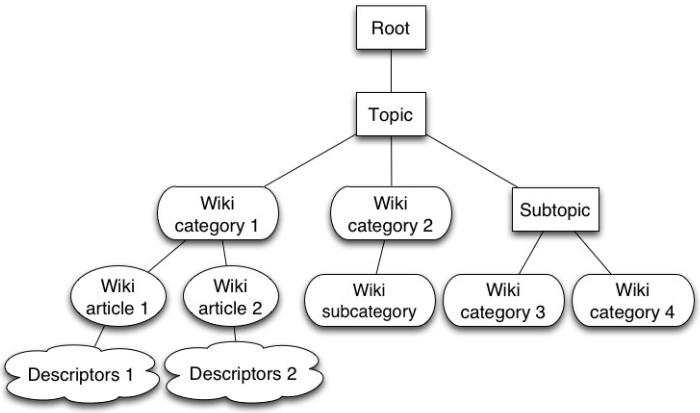

Fig. 1: The refining scheme. Initial taxonomy topics are in rectangles, the Wikipedia categories and subcategories are in rounded rectangles, the Wikipedia articles are in the ellipses, and the leaf descriptors are in the clouds.

that correspond to our taxonomy topics they should be subdivisions of the topics. Next we check, whether the assigned category should be further subdivided according to the structure of the category tree. If not, the underlying categories are again assigned to taxonomy topics. Since almost every Wikipedia category contains several articles, the titles of these articles become the leaves of the refined taxonomy. Finally, we extract keywords representing the content of each Wikipedia article. These keywords are used then as the leaves descriptors.

On the whole, the refined taxonomy should follow the gold standard of ACM CCS taxonomy: every branch of taxonomy is approximately of the same depth and has the same number of children. Therefore, each topic is refined by one ore more levels of Wikipedia categories and articles, placed on the last level as leaves.

The structure of Wikipedia is rather messy. The category tree is not a tree, but a graph, since it gets cycles within it. Some categories or articles have no semantic relation to their parental categories, the more so with regard to grandparent categories. Some categories semantically have nothing to do with their parental categories; the more so with regard to the grandparent categories. One of the explanations of this phenomenon is given in[8]: Wikipedia users tend to assign to article or subcategory as many categories as possible. For example, the category Killed accidentally lies under the category Randomness, which doesnt make sense at all.

To be able to use the Wikipedia data for taxonomy refinement, we need maintain only those essential parts that are semantically connected and have the tree-like structure. Hence the category tree should be first cleared from all irrelevant subcategories and articles: the clearing action is a step which is necessary to obtain meaningful results. Here are the main steps of our approach to taxonomy refining:

1. Specify the domain of taxonomy to be refined and set the frame of taxonomy manually.
2. Download from the Wikipedia the category tree and articles from the domain under consideration.
3. Clear the category subtree of irrelevant articles.
4. Clear the category subtree of irrelevant subcategories.
5. Assign the Wikipedia categories to the taxonomy topics.
6. Form the intermediate levels of the taxonomy
7. Use Wikipedia articles in each added category node as the leaves.
8. Extract the keywords from Wikipedia articles and use them as leaf descriptors.

Let us describe these steps in more detail using both the Probability theory and mathematical statistics (PTMS) and Numerical analysis (NA) examples.

2.1. Specify the domain of taxonomy

PTMS or NA. See tables 1 and 2 for the frames of both taxonomies.

2.2. Download from the Wikipedia the category tree and articles

Download from the Wikipedia the category trees, rooted at Probability theory and mathematical statistics and Numerical analysis and all underlying articles.

Table 3: The total number of subcategories and articles in PTMS and NA categories in Russian Wikipedia (accessed in August, 2013)

Domain	#categories	#articles
PTMS	54	928
NA	91	1340

2.3. Clear the category subtree of irrelevant articles

We consider that an article in irrelevant to the domain under consideration, if

A The relevance between the article title to the text of the articles is low;

B The relevance between the parent category title to the text of the articles is low.

The first condition allows us to filter out stubs (short unfinished article or article templates). According to the second condition we remove those articles that unlikely have something to do with the parent categories. The relevance between the title of the parent category and the article is estimated with the string-to-text relevance measure, which follows from the annotated suffix tree (AST) method (described later). It expresses conditional probability of string symbols to occur, averaged over matching fragments in suffix trees, representing a text. It ranges from 0 to 1. The smaller it is, the less is the chance the string (the title of the parent category) is relevant to the text (the article). We set up of the relevance threshold at the value of 0.2.

Table 4: Examples of irrelevant articles according to condition B

Domain	Relevance value	Category	Article
PTMS	0.9144	Probability theory	Random matrix
NA	0.1948	Regression analysis	ROC curve

2.4. Clear the category subtree of irrelevant subcategories

We declare that a subcategory is irrelevant if the similarity between its parent category title and the text obtained by merging all the articles in the subcategory is low. The relevance threshold here is set up again at the value of 0.2. This approach may fail if the subcategory does not contain any articles, but is further divided in several subcategories, so there is nothing to merge.

Table 5: Examples of irrelevant categories

Domain	Relevance value	Category	Article
PTMS	0.1515	Machine learning	Optimization theory
NA	0.0962	Noise	Noise reduction

2.5. Assign the Wikipedia categories to the taxonomy topics

After clearing the category tree from irrelevant categories and articles, we assign each of the remaining Wikipedia categories to a corresponding topic in the current fragment of taxonomy using, again, the AST relevance between the taxonomy topics and the categories represented by all their articles merged.

Table 6: Examples of category to topic assignment

Domain	Relevance value	Category	Article
PTMS	0.5323	Probability theory	Bayesian statistics
NA	0.6210	Algorithms for numerical problem solving	Algorithms for solving SLE

2.6. Form the intermediate levels of the taxonomy

The categories, which are more relevant to parent categories, than to taxonomy topics, remain as intermediate levels in the new taxonomy.

Table 7: Examples of categories, that form intermediate levels

Domain	Relevance value	Category	Article
PTMS	0.4813	Random processes	Markov processes
NA	0.3103	Algorithm efficiency	TransformersTransducers

2.7. Use Wikipedia articles in each added category node as the leaves

If a Wikipedia category is assigned to a taxonomy topic, all the articles left in it after clearing procedures are put as new leaves descending from the topic. For example, in the intermediate category Discrete distributions are put the following leaves:

- Binominal distribution
- Geometric distribution
- Poisson distribution
- Bernoulli distribution
- etc.

2.8. Extract the keywords from Wikipedia articles and use them as leaf descriptors

A leaf taxonomy topic can be assigned with a set of phrases falling in it, as is the case of ACM-CCS. To extract keywords and key-phrases, we dont employ any sophisticated techniques and take the most frequent nouns and the most frequent collocations, respectively. Of course, a key phrase is looked for as a grammar pattern, such as adjective + noun or noun + noun.

3. AST method

The suffix tree is a data structure used for storing of and searching for symbolic strings and their fragments[6]. In a sense, the suffix tree model is an alternative to the Vector Space Model (VSM), arguably, the most popular model for text representation[19]. When the suffix tree representation is used, the text is considered as a set of strings, where a string may be any semantically significant part of the text, like a word, a phrase or even a whole sentence. An annotated suffix tree (AST) is a suffix tree whose nodes (not edges!) are annotated by the frequencies of the strings

fragments. An algorithm for the construction and the usage of AST for spam-filtering is described in[11], and some other applications in[2,3] .

In our computations, we consider a Wikipedia article to be a set of three-word strings. The titles of the Wikipedia categories and articles are also considered as strings in the set. To estimate the relevance of a standalone string to a collection of strings, we build an AST for the set of strings and then find all the matches between the AST and fragments of the given string. For every match we compute the score as the average frequency of a symbol in it related to the frequency of its prefix. Then the total score is calculated as the average score of all the matches. Obviously, the final value has a flavor of the conditional probability and lies between 0 and 1. In contrast to similarity measures used in[2,3,11], this one has a natural interpretation and, moreover, does not depend on the text length explicitly, and, as our experiments tell us, implicitly.

4. Results

For the PTMS taxonomy the resulting tree has 6 levels, with its depth varying from 4 to 6. At the clearing stage 20 categories and 108 articles were removed from the Wikipedia category tree. The resulting taxonomy of numerical analysis is similar shape: it has 8 levels, the depth varies from 4 to 8. Again at the clearing stage 11 categories and 30 articles were removed. There are several problems with both obtained taxonomy trees. First, the position of the topic Decision Trees is misleading. According to our method, this topic should be placed under Mathematical statistics and be, thus, a sibling of the Machine Learning topic. The reason is the low relevance of the string Machine learning to any of the four articles in the Decision tree category. Second, the category Transformers/Transducers ([Preobrazoveteli] in Russian), which is relevant to the parent category Algorithm efficiency is further subdivided in Piezoelectrics, Power sources, Sound senders and detectors, which has nothing to do with algorithms, because of the double meaning the category title has. Third, both taxonomies stuffed with articles, describing some personalities, like Probability theorists or MIPT Lecturers. Hence some more clearing procedures, including filtering articles according to their types should be developed.

To refine a taxonomy at a given topic, the AST method works five times:

- Twice to clear the Wikipedia category tree of irrelevant articles;
- To clear the category tree of irrelevant categories;
- To relate taxonomy topics to Wikipedia categories.
- To distinguish between categories to be assigned to taxonomy topics and categories to remain as intermediate levels

In the first three cases an irrelevance threshold for the article or category title to text should be specified. Our experiments show that the threshold of 0.2, which amounts to 1/3 of the maximum value, works well.

5. Conclusion

The approach of automated refinement is part of a two-step approach to taxonomy building. First step: an expert sets a frame of the taxonomy. Second step: this frame is refined topic-by-topic until an appropriate level of granularity is reached. This approach allows protecting the taxonomy being built from noise, such as irrelevant or too detailed topics. Wikipedia is a good source for new taxonomy topics, because it contains both structured (the category tree) and unstructured (articles) data.

The presented implementation of the approach, by using an AST based relevance estimates, bears both positive and negative effects. The positive relates to the independence on the language and its grammar; and the negative, with the lack of tools for capturing synonymy and near-synonymy. This method is of little help when there is no word by word coincidence, which should be one of the main subjects for the further developments. The other direction for development include more precise Wikipedia preprocessing and analysis, such as distinguishing between different types of articles.

References

1. ACM Computing Classification System (ACM CCS), 1998, available at: http://www.acm.org/about/class/ccs98-html .
2. Chernyak E.L., Chugunova O.N., Mirkin B.G., Annotated suffix tree method for measuring degree of string to text belongingness, Business Informatics, 2012. Vol. 21, no.3, pp. 31-41 (in Russian).
3. Chernyak E.L., Chugunova O.N., Askarova J.A., Nascimento S., Mirkin B.G., Abstracting concepts from text documents by using an ontology, in Proceedings of the 1st International Workshop on Concept Discovery in Unstructured Data. 2011, pp. 21-31.
4. Cui C.,Lu Q., Li W., Chen Y. ,Mining Concepts from Wikipedia for Ontology Construction, in Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009, Vol. 3, pp.287-290.
5. Grau B.C., Parsia B., Sirin E. Working with Multiple Ontologies on the Semantic Web, in Proceedings of the 3d International Semantic Web Conference, 2004, pp. 620-634.
6. Gusfield D., Algorithms on Strings, Trees, and Sequences, Cambridge University Press, 1997.
7. Higher Attestation Commission of RF Reference, 2009, available at: http://vak.ed.gov.ru/ru/help_desk/
8. Kittur A., Chi E.H., Suh B. Whats in Wikipedia? Mapping topics and conflict using socially annotated category structure, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2009, pp. 1509-1512.
9. Liu X., Song, Y., Liu S., Wang H. Automatic Taxonomy Construction from Keywords, in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 1433-1441.
10. Loukachevitch N.V., Thesauri in information retrieval tasks, MSU, Moscow, 2011 (in Russian).
11. Pampapathi R., Mirkin B., Levene M., A suffix tree approach to anti-spam email filtering, Machine Learning, 2006, Vol. 65, no.1, pp. 309-338.
12. Ponzetto S.P., Strube M. Deriving a Large Scale Taxonomy from Wikipedia, in Proceedings of AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2007, pp. 78-85.
13. Robinson P.N., Bauer, S., Introduction to Bio-Ontologies, Chapman & Hall/CRC, USA, 2011.
14. Sadikov E., Madhavan J., Wang L., Halevy A.Y., Clustering query refinements by user intent, in Proceedings of the 19th International Conference on World Wide Web, 2008 pp. 841-850.
15. Taxonomy of Abstracting Journal Mathematics , VINITI. Available at: http://www.viniti.ru/russian/math/files/271.htm , 1999, (in Russian).
16. Van Hage W.R., Katrenko S., Schreiber G., A Method to Combine Linguistic Ontology-Mapping Techniques, in Proceedings of 4th International Semantic Web Conference, 2005, pp. 34-39.
17. White R.W., Bennett P.N., Dumais S.T. Predicting short-term interests using activity-based search contexts, in Proceedings of 19th ACM conference on Information and Knowledge Management, 2010, pp. 1009-1018.
18. Wu F., Weld. D. Automatically refining Wikipedia Infobox Ontology, in Proceedings of the 17th International World Wide Web Conference, 2008, pp. 635-645.
19. Zamir O, Etzioni. O. Web document clustering: A feasibility demonstration, in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 46-54.
20. Zirn C., Nastase V., Struve M., Distinguishing between Instances and Classes in the Wikipedia Taxonomy, in Proceedings of 5th European Semantic Web Conference, 2008, pp. 376-387.