АВТОМАТИЧЕСКОЕ ДОСТРАИВАНИЕ ТАКСОНОМИИ НА РУССКОМ ЯЗЫКЕ НА ОСНОВЕ РЕСУРСОВ ВИКИПЕДИИ

Черняк Е. Л. (echernyak@hse.ru),
Миркин Б. Г. (bmirkin@hse.ru)

Отделение прикладной математики и информатики,
НИУ Высшая Школа Экономики, Москва, Россия

Ключевые слова: достраивание таксономии, близость между строкой и текстом, использование Википедии, суффиксные деревья

COMPUTATIONAL REFINING OF A RUSSIAN-LANGUAGE TAXONOMY USING WIKIPEDIA

Chernyak E. L. (echernyak@hse.ru),
Mirkin B. G. (bmirkin@hse.ru)

Department of Applied Mathematics and Informatics, National Research University, Higher School of Economics, Moscow, Russia

A two-step approach to devising a hierarchical taxonomy of a domain is outlined. As the first step, a coarse “high-rank” taxonomy frame is built manually using the materials of the government and other representative sites. As the second step, the frame is refined topic-by-topic using the Russian Wikipedia category tree and articles filtered of “noise”. A topic-to-text similarity score, based on annotated suffix trees, is used throughout. The method consists of three main stages: 1) clearing Wikipedia data of noise, such as irrelevant articles and categories; 2) refining the taxonomy frame with the remaining relevant Wikipedia categories and articles; 3) extracting key words and phrases from Wikipedia articles. Also, a set of so-called descriptors is assigned to every leaf; these are phrases explaining aspects of the leaf topic. In contrast to many existing taxonomies, our resulting taxonomy is balanced so that all the branches are of similar depths and similar numbers of leaves. The method is illustrated by its application to a mathematics domain, “Probability theory and mathematical statistics”.

Keywords: taxonomy refinement, string-to-text similarity, 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”. Automating the process of taxonomy building is important for further progress of computational text processing and 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 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 [5]. 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 “coarse”, 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 topic-to-text similarity score, based on annotated suffix trees, is used throughout. Our method for refining of a taxonomy leaf, after the relevant materials from Wikipedia are downloaded, involves removing “irrelevant” subjects and articles. The method is illustrated by its application to a mathematics domain, “Probability theory and mathematical statistics” (in Russian), which highlights both advantages and drawbacks of the method.

This application is relevant to our work on using taxonomies for computational visualization and interpretation of published paper abstracts and university 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; it covers just two layers of the body of science. Two or three more layers can be derived from the so-called HAC 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 us 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], one of the most advanced domain taxonomies, 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 starting any refinement steps 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 [4]. 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 [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. Reference [5] lists these advantages of using Wikipedia in taxonomy building:

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

In papers [12,16,18] different approaches for constructing or refining ontologies and taxonomies by using Wikipedia article data are presented. In [12] the Wikipedia articles, in [20], the Wikipedia category tree, and in [18], 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, and moreover, we score the extent of relevance of those to the parental category. 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.

2. Our approach to taxonomy refinement using Wikipedia

We specify the taxonomy frame manually by extracting basic topics from the publicly available instruction materials of the Higher Attestation Commission of Russia [7]. The HAC materials are reflected in a three-level rooted tree of the main topics of probability theory and mathematical statistics (see Table 1).

Table 1. HAC based "Probability theory and mathematical statistics" taxonomy frame

Probability theory and mathematical statistics		
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

We use the corresponding Wikipedia category, that is, "The Probability Theory and Mathematical Statistics", as the only source for new topics. Luckily, the topic of our interest is a category in Wikipedia, so there is no need to address any other categories. For our purposes, it is useful to distinguish between two Wikipedia data types:

1. The hierarchical structure of Wikipedia category tree
2. The collection of unstructured Wikipedia articles.

Hereafter we are going to use the Wikipedia category tree for extending our taxonomy tree, whereas the articles are used as the source of keywords. We try to assign Wikipedia categories and the underlying subcategories to every taxonomy topic of the first and second levels. First, we find those Wikipedia categories that correspond to our taxonomy topics — they should be subdivisions of the topics. Each subdivision is further divided according to the Wikipedia articles in that, so that the titles of the articles are leaves of the final taxonomy tree. Then, we extract keywords representing the content of each Wikipedia article. These keywords are used then as leaf descriptors.

Therefore, each topic is refined in a two-level subtree, which consists of a Wikipedia category and corresponding Wikipedia articles.

Unfortunately, the structure of the Russian Wikipedia categories is rather noisy. Some categories semantically have nothing to do with their parental categories. For example, in the Russian Wikipedia category tree, the Optimization category lies under the Machine learning category, which itself falls in the Mathematical Statistics category (accessed December 2012). Moreover, the category tree in some places loses its tree-like format and gets cycles within it. One of the explanations of this phenomenon is given in [8]: Wikipedia users' passion to category assignment.

To be used for taxonomy refining, the relevant part of the category tree should be first cleared from all irrelevant subcategories and articles: the clearing action appears to be necessary for obtaining meaningful results.

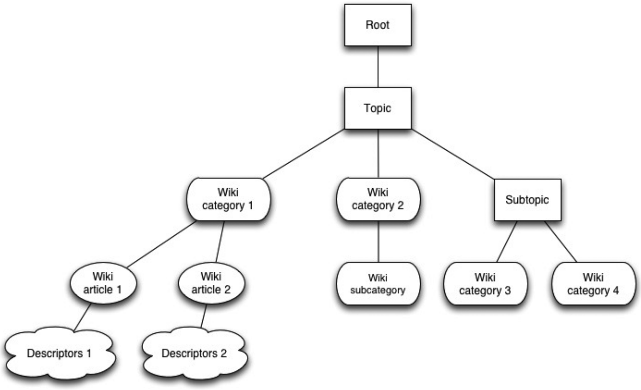

Fig. 1. Our refining scheme: Initial taxonomy topics are in rectangles, the Wikipedia categories and subcategories are in rounded rectangles, the Wikipedia articles are in the ellipses, and the leaf descriptors are in the clouds

Here are the main steps of our approach to taxonomy refining:

1. Specify a taxonomy topic to be refined.
2. Download the related Wikipedia category subtree and the articles from the taxonomy topic under consideration.
3. Clear the category subtree of irrelevant articles.
4. Clear the category subtree of irrelevant subcategories.
5. Extend the taxonomy tree in the specified topic node with the cleared Wikipedia subtree.
6. Put Wikipedia articles in each added category node as the leaves.
7. Extract keywords from Wikipedia articles and use them as leaf descriptors.

Let us illustrate these steps using the following manually made example.

2.1. Specifying taxonomy topic:

Probability Theory and Mathematical Statistics (PTMS).

2.2. Downloading the contents of the topic according to Wikipedia

Download from Wikipedia the subtree of concepts rooted at PTMS. There were 640 Wikipedia articles assigned in 48 categories.

2.3. Clearing the category tree of irrelevant articles

Some of the nodes in the downloaded tree have obviously nothing to do with the PTMS subjects, such as “Software optimization” and “Natural language toolkit”. We consider that an article is irrelevant if the similarity between the parent category title ant the text of the article is low; setting a threshold value is described in section 3. The similarity value follows from the annotated suffix tree (described later) and ranges from 0 to 1. It expresses the average level of conditional probability of a symbol in a string to appear after the string’s prefix.

2.4. Clearing the category tree 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; setting a threshold value is described in section 3. Unfortunately, this approach may fail sometimes. For example, the Decision Tree subcategory is irrelevant to the Machine Learning according to our rule, which is obviously wrong. The cause: none of the four articles in the category Decision Tree contain phrase "Machine Learning" or any of its substrings.

2.5. Extending the taxonomy tree by Wikipedia categories

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 similarity between the taxonomy topics and the categories represented by all their articles merged.

2.6. Putting Wikipedia articles as the taxonomy leaves to the taxonomy tree

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.

2.7. Extracting keywords from Wikipedia articles and using them as descriptors to leaves

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 don't 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, that is, any semantically significant parts of 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 similarity between a standalone string and 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 show, implicitly. To specify an "irrelevance" threshold for the similarity between a category and a text, we take the threshold of 0.2, which amounts to 1/3 of the maximum similarity value and, in our experiments, works well.

4. Results

For the taxonomy in Table 1 the resulting taxonomy tree has 7 levels, with its depth varying from 4 to 7. A fragment of the tree is presented on Figure 2. At the clearing steps a hundred irrelevant articles and two irrelevant categories were removed from the Wikipedia category subtree. Some of the taxonomy topics remain untouched as, for example, "Methods for Statistical Analysis and Inference".

There is a problem with the obtained taxonomy tree: the position of the topic "Decision Trees". According to our method, this topic should be placed under "Multivariate Statistics and Data Analysis" and be, thus, a sibling of the "Machine Learning" topic. Moreover, as mentioned above, the "Decision Trees" has a very low similarity to "Machine Learning".

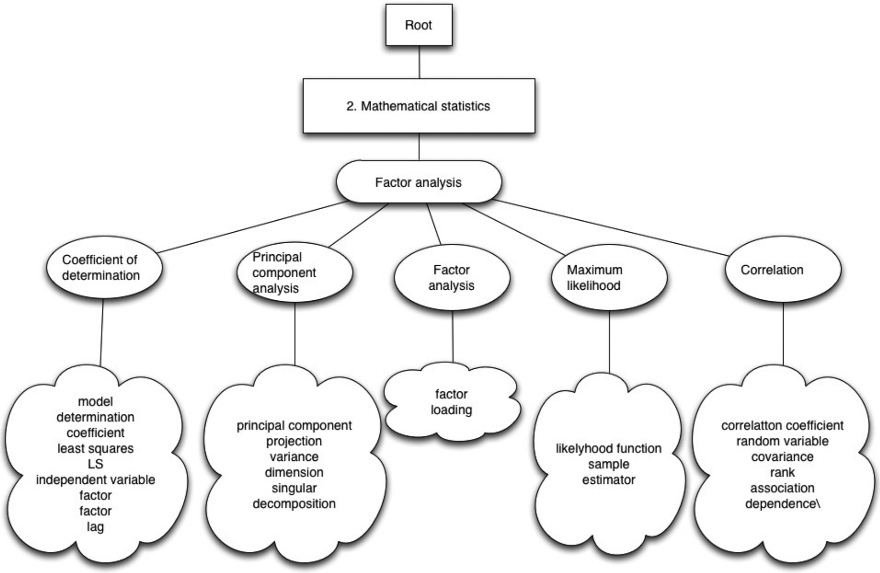

Fig. 2. A fragment of the refined taxonomy tree: the "Factor Analysis" branch

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

1. Clear the Wikipedia category subtree of irrelevant articles;
2. Clear the category subtree of irrelevant categories;
3. Relate taxonomy topics to Wikipedia categories.

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 similarity 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.

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.* (2012), Annotated suffix tree method for measuring degree of string to text belongingness [Metod annotirovannogo suffiksnogo dereva dlja otsenki stepeni vhozhdenija strok v tekstovie dokument], Biznes-Informatika [Business Informatics], no.3, pp. 31–41.
3. *Chernyak E. L., Chugunova O. N., Askarova J. A., Nascimento S., Mirkin B. G.* Abstracting concepts from text documents by using an ontology. Proceedings of the 1st International Workshop on Concept Discovery in Unstructured Data. Moscow, 2011, pp. 21–31.
4. *Grau B. C., Parsia B., Sirin E.* Working with Multiple Ontologies on the Semantic Web. In Proceedings of the 3d International Semantic Web Conference, Hiroshima, Japan 2004, pp. 620–634.
5. *Grineva M., Grinev M., Lizorkin D.* (2009), Text documents analysis for thematically grouped key terms extraction [Analis tekstovih dokumentov dlja isvlechenenija tematicheski sgruppirovannih kljuchevih terminov], in Trudy Instituta sistemnogo programmirovanija RAN [Works of Institute for System Programming of the RAS], Institute for System Programming, pp. 155–156.
6. *Gusfield D.* (1997), Algorithms on Strings, Trees, and Sequences, Cambridge University Press.
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.* What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, USA, 2009, pp. 1509–1512.
9. *Liu X., Song, Y., Liu S., Wang H.* Automatic Taxonomy Construction from Keywords. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, New York, 2012, pp. 1433–1441.
10. *Loukachevitch N. V.* (2011), Tezaurusy v zadachah informatsionnogo poiska [Thesauri in information retrieval tasks], MSU, Moscow.
11. *Pampapathi R., Mirkin B., Levene M.* (2006), A suffix tree approach to anti-spam email filtering, Machine Learning, Vol. 65(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.* (2011), Introduction to Bio-Ontologies, Chapman & Hall/CRC, USA.
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, New York, USA, 2008, pp. 841–850.
15. *Taxonomy of Abstracting Journal "Mathematics"* (1999), VINITI.
    Available at: http://www.viniti.ru/russian/math/files/271.htm
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, Galway, Ireland, 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, Toronto, Canada, 2010, pp. 1009–1018.
18. *Wu F., Weld D.* Automatically refining Wikipedia Infobox Ontology. In Proceedings of the 17th International World Wide Web Conference, Beijing, China, 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, Seattle, USA, 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, Tenerife, Spain, 2008, pp. 376–387.