Categorizing Learning Objects Based On Wikipedia as Substitute Corpus

Marek Meyer[1,2], Christoph Rensing[2], and Ralf Steinmetz[2]

[1] SAP AG, SAP Research CEC Darmstadt, Bleichstr. 8, 64283, Germany,
marek.meyer@sap.com,
[2] KOM Multimedia Communications Lab, Technische Universität Darmstadt,
Merckstr. 25, 64832 Darmstadt, Germany,
{rensing, steinmetz}@kom.tu-darmstadt.de

Abstract. As metadata is often not sufficiently provided by authors of Learning Resources, automatic metadata generation methods are used to create metadata afterwards. One kind of metadata is categorization, particularly the partition of Learning Resources into distinct subject categories. A disadvantage of state-of-the-art categorization methods is that they require corpora of sample Learning Resources. Unfortunately, large corpora of well-labeled Learning Resources are rare. This paper presents a new approach for the task of subject categorization of Learning Resources. Instead of using typical Learning Resources, the free encyclopedia Wikipedia is applied as training corpus. The approach presented in this paper is to apply the k-Nearest-Neighbors method for comparing a Learning Resource to Wikipedia articles. Different parameters have been evaluated regarding their impact on the categorization performance.

1 Introduction

One of the success factors of e-Learning is the availability and reusability of existing Learning Resources. Learning Object Repositories (LOR) are used to collect and disseminate Learning Resources with others. But re-use of a Learning Resource requires not only availability but also that it can be found. Typically, retrieval of Learning Resources is based on metadata. Metadata records contain information about the contents of Learning Resources.

In practice, authors often do not provide enough metadata. Thus, it is necessary to generate or supplement metadata records after they have been uploaded to a repository. This a posteriori generation can be performed either manually by humans or automatically by algorithms. For large LORs only automatic metadata generation is feasible.

This paper focuses on the generation of topical metadata. A new approach for categorizing Learning Objects into subject categories is presented. The free encyclopedia Wikipedia is used as a corpus for classification methods.

The next section discusses related work concerning automatic metadata generation. Section 3 introduces the idea of using Wikipedia articles as substitute corpus for classification. The actual experiment is presented in section 4 and section 5 illustrates the evaluation results.

2 Automatic Metadata Generation

Metadata generation is a field of research that has been heavily worked on in the recent years. There are many approaches for metadata generation for documents in general [1] and for Learning Resources in particular [2]. Metadata generation methods can be classified by the type of metadata to generate, by the sources that are used, by the required prerequisites and the applied methods. A framework for automatic metadata generation has been proposed by Cardinaels et al. [3].

Possible target metadata types are for example content-related metadata (such as title, keywords and categories), process-related metadata (author, creation date, version) or didactical metadata (learning objective, target group, difficulty, activity level). Sources for metadata generation strongly depend on the target metadata types. Content-related metadata requires to analyze the contents of a document, whereas process metadata, such as author and creation date can be obtained from the authoring environment [4].

The focus of this paper is the generation of topical metadata, in particular categorization by subject. Subject categorization can be divided into domain-restricted and open domain categorization. Domain-restricted category systems are used for repositories or communities with a limited scope of topics (e.g. only Learning Resources about Mathematics and Computer Science), whereas open domain category systems try to cover any possible topic (e.g. the category systems of public libraries).

A typical solution to classification tasks is to apply machine learning methods. Machine learning methods train classifiers with a set of sample objects, for which the correct category assignment is known. Exemplary methods are support vector machines, Bayes classifiers, k-Nearest-Neighbors classifiers or decision tree learners [5]. Using machine learning methods for classifying Learning Resources has one drawback. These methods require a corpus of training samples; for each category several samples are needed for achieving good results. But a large enough training corpus is most often not available when establishing new repositories.

3 Wikipedia as Substitute Corpus

As described in the previous section, the application of machine learning methods for classifying Learning Resources requires to have a test corpus at hand. For each supported category there have to be several sample objects in order to achieve a good classification performance. In the area of Learning Object Repositories, an adequate corpus is often not available. In particular for open domain LORs – repositories that accept Learning Resources about any topic – a training corpus is missing. And even for restricted domain repositories the manual creation of a training corpus causes a high effort.

The underlying idea of this paper is the usage of an encyclopedia as a substitute corpus. The free encyclopedia Wikipedia [6] has been chosen as data source, because a database dump is downloadable for free. In an earlier experiment, the suitability of Wikipedia has been tested [7]. In addition, Gabrilovich

and Markovitch have shown that Wikipedia can also be used for improving classification in other areas than e-Learning [8]. The basic approach is to transform all Wikpedia articles into a word vector representation. Learning Resources are also mapped to word vectors and compared to the article vectors; articles, which are very close to the Learning Resource vector are assumed to cover a similar topic.

Categorization requires a category system from which the labels for each object are taken. Many repositories already provide a category system (or catalog). Wikipedia also has a category system, into which all articles are organized. In general, the categories present in Wikpedia are very comprehensive and well-balanced. The categories of Wikipedia are a consensus of many subject matter experts. Thus, for open-domain repositories the Wikipedia categories may be adopted as category system. In total the German Wikipedia contains about 41,000 categories, the English version has even more. The Wikipedia category system is modeled as a directed acyclic graph. Each category contains links to one or multiple more generic categories.

The basic idea of using Wikipedia articles as corpus is to use all articles, which belong to a particular category, as training samples for that category. Different machine learning algorithms may be used in principle. However, the total amount of categories and articles imposes special requirements on the applied methods. When all categories are used, memory consumption and calculation complexity become limiting factors for the choice of classifiers. Because of these limits, the k-Nearest-Neighbors (kNN) approach has been chosen for the experiments, which are described in this paper. A kNN classifier compares an object to classify in a vector space model to the training objects. The k nearest objects are used to determine the category of the input vector.

4 Experiment Setup

An experiment has been performed for evaluating the performance of the described approach. This experiment is supposed to result in a quantitative evaluation, whereas [7] only provided a qualitative statement. Another expected outcome is an evaluation of different parameters of the method.

First of all, classification can be performed either as single-label or multi-label assignment. Single-label classification requires that each object belongs to exactly one category. Multi-label classification allows objects to belong to several categories. The present paper applies single-label classification.

The hierarchical structure of the category system makes classification and evaluation a complex task [9]. There are different approaches for hierarchical classification and evaluation. The classifier may either ignore the hierarchy (flat classification) or involve the structure into the decision. A typical approach for hierarchy-aware classification is to classify top-down one hierarchy level after the other. At each hierarchy level two classifiers are employed: a local classifier decides on a given hierarchy level to which subcategory an object belongs. A second classifier determines if the object belongs to a subcategory at all or if

recursion ends [10]. The experiments in this paper are performed with both the flat classification and a variation of the hierarchical approach. A method called hierarchical propagation (HP) starts with flat classification but also recursively contributes to the ranking of more generic categories. A propagation rate value controls how strong this contribution is.

A set of 100 Learning Resources from the k-Med project [11] serve as test set. These Learning Resources are web-based training courses from the area of medical science. All 100 courses were manually classified for having reference labels. Each Learning Resource is assigned to exactly one category.

The implemented classification method is based on the k-Nearest-Neigbors approach. The text of each Learning Resource to classify is transformed into a word vector. This word vector is compared to the word vectors of all Wikipedia articles. Different similarity measures for vector spaces are known, which may produce differing results. Hence, different similarity measures have to be evaluated for their usefulness. According to the kNN method the k most similar articles are used for determining the classification of a Learning Resource.

The experiment has two goals. The first goal is to test the feasibility of the classification approach. A second goal is to determine how the method can be optimized. For the second goal, four parameters are varied and optimized. The four parameters are as follows.

- Feature selection (used words)
- Similarity measures in word vector space
- k-Value for k-Nearest-Neighbors method
- Method for mapping articles to a category

The first parameter is feature selection. Each Learning Resource contains many different words, which can be used for classification. However, some very frequent words, such as "a", "the" or "are" have no or even negative influence on the classification performance. Some very rare terms could also compromise the results. For the experiments seven different word lists are used. One contains all stemmed words that occur in any Wikipedia article. The remaining six word lists are combinations of high frequency pruning and low frequency pruning. For low frequencies either all terms are used or only those, which occur in at least three articles. High frequencies are cut above document frequencies of 200,000, 100,000 and 50,000.

As second parameter different similarity measures are applied for comparing documents in vector space. The four similarity measures are cosine similarity, Jaccard similarity coefficient, binary Dice's coefficient and overlap coefficient. The cosine similarity is based on the TF-IDF weighting, whereas the other three measures work on binary values [5].

A further parameter is the k value for the kNN method. k is instantiated with the values 1, 3, 5, 10, 20 and 30.

The last parameter is the applied classification method that maps from a set of articles to the most probable category. Two different methods are used: flat classification and hierarchical propagation (HP). Flat classification ignores the

hierarchical structure of the category system and simply determines the category, to which most of the k articles immediately belong. Hierarchical propagation, on the other hand, also considers the more generic categories of an article. Each article of the kNN set recursively propagates the occurrence to its superior categories multiplied with a propagation rate p. Four different values for p are tested: 0.2, 0.3, 0.5 and 0.7.

The performance of a classifier can be measured in different ways. The most common measures in machine learning are accuracy, precision and recall. Accuracy is an overall value that indicates how many objects are assigned to the correct category. Precision and recall are calculated separately for each category. The precision value tells how many of the objects classified for a particular category are correct. Recall indicates how many objects that are known to belong to a category are correctly classified to that category by the classifier.

In the case of hierarchical classes, these values are not adequate, as they do no take into account generalization and specialization of categories, as well as sibling relationships [10]. Consider for example a Learning Resource about *passenger cars*. If the categorization algorithm proposes the category *cars*, this not the expected result, but on the other hand not completely wrong. Thus, the performance evaluation of this paper is based on two different measures. The first one is a simple accuracy value, which is the percentage of Learning Resources which have been classified exactly into the correct category. A second measure is the so called average category link distance (CLD). For each Learning Resource the distance between the classified category and the nominal category is calculated; the distance is defined here as number of edges in the category graph. The average of all link distances is calculated for all parameter combinations. The CLD measure has been chosen because it is easy to implement. More sophisticated measures are, for instance, Resnik's information content [12] or the category similarity of Sun and Lim [10].

5 Experimental Evaluation

The experiment was performed as described in the previous section. For all combinations of the four parameters the resulting classification were evaluated according to two basic measures: the average category link distance (CLD) and direct accuracy (percentage of objects with CLD of 0). In some cases a classifier could not classify a Learning Resource, for instance because the determined articles belong to no category. In this case, a category link distance could not be calculated. Therefore, a derived measure has been introduced: a second CLD value (CLD 2) is calculated only of tangible categories.

A first matrix analyzes which combination of word list (feature selection) and similarity measure performs best. Table 1 presents the performance values for all combinations of word lists and similarity measures. The performance values aggregate the results of all combinations of the remaining two parameters, listing the minimum of CLD values and the maximum of accuracy values. Obviously, the cosine measure significantly outperforms the other similarity measures, no

matter which word list is used. Regarding the effect of different word lists, the low frequency pruning predominantly improves the results. Cutting high frequency terms also causes improvements. But the optimal word lists differs between the similarity measures. In any case, pruning terms with a document frequency above 200,000 pays off.

Table 1. Evalution for similarity measures and feature selection.

simil. measure	performance indicator	all terms	1-50k	1-100k	1-200k	3-50k	3-100k	3-200k
cosine	min CLD	1.24	1.10	1.14	1.19	1.09	1.05	1.08
	min CLD 2	1.19	1.10	1.14	1.19	1.09	1.05	1.08
	max Accuracy	57%	60%	59%	58%	61%	62%	62%
jaccard	min CLD	1.76	1.48	1.65	1.75	1.49	1.66	1.66
	min CLD 2	1.76	1.48	1.65	1.75	1.49	1.66	1.66
	max Accuracy	35%	43%	40%	37%	43%	39%	37%
dice	min CLD	1.76	1.48	1.65	1.75	1.49	1.66	1.66
	min CLD 2	1.76	1.48	1.65	1.75	1.49	1.66	1.66
	max Accuracy	35%	43%	40%	37%	43%	39%	37%
overlap	min CLD	–	–	–	–	–	–	–
	min CLD 2	3.79	3.62	3.67	3.74	3.45	3.62	3.55
	max Accuracy	2%	3%	7%	4%	3%	6%	3%

As Table 1 has shown, the combination of the cosine similarity measure and a word list with document frequencies between three and 100,000 provide best results. Table 2 displays the performance of the different category mapping methods, given that similarity measure and word list are set to these values. The results are again aggregated over all values of k. The numbers of Table 2 indicate that the flat classification approach performs best.

Table 2. Evalution for similarity measures and feature selection.

perf. indicator	flat	HP-0.2	HP-0.3	HP-0.5	HP-0.7
min CLD	1.05	1.22	1.27	2.84	3.75
min CLD 2	1.05	1.17	1.21	1.86	3.39
max Accuracy	62%	56%	54%	33%	4%

Finally, Table 3 shows the performance of the kNN classifier for different values of k. The optimal parameters from the previous tables (cosine similarity measure, 3-100k word list and flat classification) are used for this evaluation. What can be learned from the table is that for values up to 10 the problem of uncategorized articles negatively impacts the results. A k of 20 performed best in the given case. However, if uncategorized articles were removed before the whole process, smaller values of k could perform better.For a larger k, the accuracy decreased again.

Table 3. Evalution for similarity measures and feature selection.

perf. indicator	k**=1**	k**=3**	k**=5**	k**=10**	k**=20**	k**=30**
CLD	–	–	–	–	1.05	1.23
CLD 2	1.59	1.51	1.54	1.11	1.05	1.23
Accuracy	24%	39%	45%	58%	62%	55%

The best result has been produced by a combination of the cosine similarity measure, a word list with document frequencies between 3 and 100,000, flat classification and a k value of 20. This combination achieved an accuracy of 62% regarding only perfect matches. In average, the determined categories were about 1.05 links in the category graph away from the nominal category. The distribution of category link distances for this parameter combination is illustrated in Fig. 5.

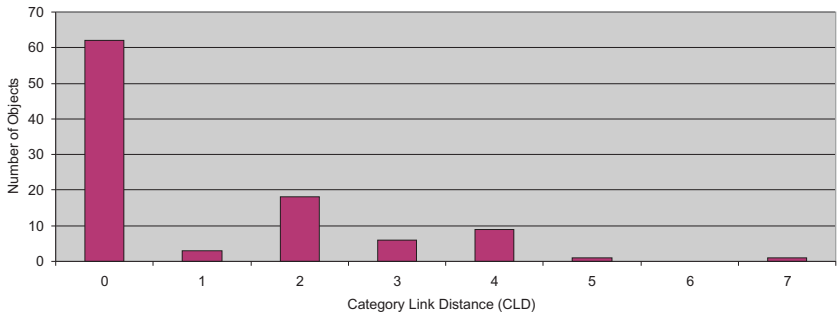

Fig. 1. Distribution of Category Link Distances.

6 Conclusions

This paper has evaluated if encyclopedic articles are suitable as a substitute corpus for the classification of Learning Resources. The k-Nearest-Neighbors method has been applied on Wikipedia articles as sample objects and the Wikipedia category system as classes. The experimental results have shown that the approach is feasible. An accuracy of 62% could be achieved with the chosen methods and parameters. Furthermore, the average deviation from the nominal category was 1.05 edges in the category graph. Different similarity measures, feature selections and further parameters have been evaluated. The cosine similarity in combination with a pruning of very rare and very frequent terms offered the best performance.

Obviously, the presented approach works only for mainly text-based Learning Resources. While for speech a transcript might be used as text source, the presented method is not applicable to images and videos.

For the future, further experiments are planned. Different classifiers beside kNN have to be evaluated. It is also assumed, that the usage of additional information from Wikipedia could increase the performance. Especially the link structure bears potential for improvements. Further classification approaches should make use of the link structure between articles. In the presented experiment, a standard kNN method has been used where each of the k obtained articles is equally weighted. A new idea is to overweight those articles that are linked to other articles in the result set. This would lessen the influence of noise.

References

1. Noufal, P.: Metadata: Automatic generation and extraction. In: 7th MANLIBNET Annual National Convention on Digital Libraries in Knowledge Management: Opportunities for Management Libraries, Indian Institute of Management Kozhikode (May 2005)
2. Bergstraesser, S.: Automatisierung der Erstellung von Metadaten. Diploma thesis, Darmstadt University of Technology (Mar 2005)
3. Cardinaels, K., Meire, M., Duval, E.: Automating metadata generation: the simple indexing interface. In: WWW '05: Proceedings of the 14th international conference on World Wide Web, New York, NY, USA, ACM Press (2005) 548–556
4. Hoermann, S., Hildebrandt, T., Rensing, C., Steinmetz, R.: ResourceCenter - A Digital Learning Object Repository with an Integrated Authoring Tool Set. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2005, Montreal, AACE (June 2005) 3453–3460
5. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1) (2002) 1–47
6. Wikipedia: The free encyclopedia. http://en.wikipedia.org (2007) [Online; last visited 22th June 2007].
7. Meyer, M., Rensing, C., Steinmetz, R.: Towards using wikipedia as a substitute corpus for topic detection and metadata generation in e-learning. In: Proceedings of the 3rd annual e-learning conference on Intelligent Interactive Learning Object Repositories (i2LOR 2006). (2006)
8. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI). (2007)
9. Dumais, S., Chen, H.: Hierarchical classification of web content. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press (2000) 256–263
10. Sun, A., Lim, E.P.: Hierarchical text classification and evaluation. In: First IEEE International Conference on Data Mining (ICDM'01), Los Alamitos, CA, USA, IEEE Computer Society (2001) 521
11. k-MED: Knowledge in Medical Education. http://www.k-med.org (last accessed: 05/2007)
12. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: IJCAI. (1995) 448–453