AISLES THROUGH THE CATEGORY FOREST
Utilising the Wikipedia Category System for Corpus Building in Machine Learning

Rüdiger Gleim, Alexander Mehler, Matthias Dehmer and Olga Pustylnikov
Text Technology Group, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany

Keywords: Social Tagging, Wikipedia, Category System, Corpus Construction.

Abstract: The Word Wide Web is a continuous challenge to machine learning. Established approaches have to be enhanced and new methods be developed in order to tackle the problem of finding and organising relevant information. It has often been motivated that semantic classifications of input documents help solving this task. But while approaches of supervised text categorisation perform quite well on genres found in written text, newly evolved genres on the web are much more demanding. In order to successfully develop approaches to web mining, respective corpora are needed. However, the composition of genre- or domain-specific web corpora is still an unsolved problem. It is time consuming to build large corpora of good quality because web pages typically lack reliable meta information. Wikipedia along with similar approaches of collaborative text production offers a way out of this dilemma. We examine how social tagging, as supported by the MediaWiki software, can be utilised as a source of corpus building. Further, we describe a representation format for social ontologies and present the *Wikipedia Category Explorer*, a tool which supports categorical views to browse through the Wikipedia and to construct domain specific corpora for machine learning.

1 INTRODUCTION

The development of the Word Wide Web has inspired new branches and continuously offers challenges in the area of information retrieval and related disciplines (e.g. topic tracking and text mining). This accounts especially to methods in supervised learning facing the peculiarities of *web* documents whose categorisation, for example, is more demanding than the one of *text* documents. One reason is that while genres of written text (e.g. scientific papers, job postings etc.) are – in terms of their structure and function – relatively stable, *web genres* (Mehler and Gleim, 2006) are still in flux and continuously emerge as the web serves more and more communication functions which, previously, were distributed over different media. Thus, in order to successfully classify instances of newly evolving web genres, a better understanding of their typical document structure, content, function and interrelations is indispensable. Solving this task is, in turn, bound to the availability of adequate and large enough corpora of natural language texts which serve as training or test data (Mitchell, 1997) for the development and evaluation of reliable approaches to web mining.

The composition of domain- or genre-specific web corpora is still an unsolved problem. The reason is that it is time consuming to build large corpora of putatively representative web pages since we generally lack trustable meta information, thus leaving it to the researcher to manually check and annotate instance pages – a task out of reach when it comes to handling highly fluctuating web data.

Knowledge communication by example of the Wikipedia project (Leuf and Cunningham, 2001) along with similar approaches of collaborative content production may offer a way out of this dilemma. By enabling users to assign wiki articles to one or more category documents, a convenient method of social classification or *social tagging* is provided. Nonetheless, utilising social ontologies and text corpora based thereon as a source of corpus building in the framework of machine learning, is highly demanding as users rarely agree on conventions of *when*

Gleim R., Mehler A., Dehmer M. and Pustylnikov O. (2007).
AISLES THROUGH THE CATEGORY FOREST - Utilising the Wikipedia Category System for Corpus Building in Machine Learning.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 142-149
DOI: 10.5220/0001267101420149

and *how* to categorise articles.

This article examines how the wiki-based approach to social tagging can be utilised to tackle the task of building domain-specific corpora in machine learning. We present an approach to extracting, representing and enhancing the category systems of wiki-based encyclopediae. This includes especially the *Wikipedia Category Explorer* (henceforth named WikiCEP), a tool which provides category system-based utilities to browse the Wikipedia as well as to select and extract domain-specific text corpora.

Section 1.1 starts with discussing related approaches to enhancing and utilising wiki-based ontologies. Section 2 gives a brief introduction to the way Wikipedia supports article categorisation and social tagging. It presents a graph representation as a starting point to analysing benefits and drawbacks of wiki category systems and, finally, proposes an enhanced graph representation model which tackles some of its flaws. Section 3 proposes an approach to how the category system of Wikipedia can be used to select articles in order to build domain specific corpora. The WikiCEP tool, which implements this approach, is presented in detail in section 4. Finally, we exemplarily sketch applications which may benefit from the concept described and give a prospect of future work.

1.1 Related Work

Shapiro (2002) has build the TouchGraph system, a graph visualisation tool of which one demo application allows to browse the link structure of small Wiki instances. In general, this tool should be adaptable to the Wikipedia, if the wiki graph is preprocessed and represented as required by TouchGraph. But facing the sheer amount of graph data, its small world topology (Mehler, 2006) and temporal variability, this is, obviously, barely manageable by a single graph viewer which works offline.

Wikipedia itself offers (i.e. by means of the MediaWiki Software[1]) a tool to browse its category system *online*. The so called CategoryTree tool[2] is directly integrated into the Wikipedia: It provides information on category documents which consist of lists of links to their respective hyponyms and hyperonyms as well as to articles being categorised by the category under consideration. Furthermore, each entry of the list of hyponyms of a category document can be expanded to a tree if the focal hyponym contains children itself. This functionality enables straightforward access to the wiki category system and can be used to manually explore the context of a category document – that is, its position within the hierarchical taxonomy of hyperonyms and hyponyms. However, practically it only allows to explore a small part of the category system. Furthermore the tree expansion of hyponyms is a bit misleading as it suggests that the underlying structure is a tree which is not the case as we will show in section 2. Thus the *on-board* means to browse through the categories misses to offer an overall picture of social ontologies which depart from classical hierarchical taxonomies as they allow users to link the same category with several putative hyperonyms, thus, making use of the expressiveness of graphs. Consequently, some kind of a GraphView is needed instead or in addition to the TreeView provided by the CategoryTree tool. This need is analysed in more detail in the following section.

2 OPERATIONALISING WIKI CATEGORY SYSTEMS

In this section we give a brief introduction to how Wikipedia allows users to classify articles according to a rich and equally flexible category system. We discuss the advantages and drawbacks of this approach by mapping the wiki category system to a graph representation model which allows to examine its features from a graph-theoretic point of view. More specifically, we propose an approach to derive a graph-like representation which is called *generalised tree* as it consists of a kernel rooted tree which is augmented by graph inducing links as the proper data structure to map wiki-based category systems. This is done by example of the German distribution of the Wikipedia[3], which, for the time being, is the second largest.

2.1 Article Categorisation In Wikipedia

Wikipedia is well known for the ease of article creation, edition and interlinking. By simply writing the name of another article into the documents' source code and putting it into doubled squared brackets an author can establish links to other articles which he or she thinks may be of interest to the reader (e.g. [[Related Article]]). That way, a complex network of highly interlinked articles has evolved sharing the peculiarities with many social networks (Newman, 2003; Zlatic et al., 2006; Mehler, 2006). So complex in fact that a mechanism was needed that could

[1] http://www.mediawiki.org/wiki/MediaWiki

[2] http://tools.wikimedia.de/~daniel/WikiSense/CategoryTree.php

[3] as extracted on 2006-08-03

improve the organisation of the contents. In order to help users to classify their contributions according to some general topic markers, a separate type of document has been introduced: The so called *category*-documents are separated from common articles by a different namespace. By linking to a category document a categorisation is expressed (e.g. by adding [[Category:Music]] to the document source the respective article is assigned to the category '*Music*'). From the readers' perspective, a categorisation is visible by means of a separate text box at the end of the document which lists the set of categories the article is assigned to. By following such a category-link, the respective category document is shown. It may contain arbitrary content as articles do but usually only offers a brief description of the category itself. The important point is that it additionally contains a list of articles which are likewise categorised by that document. Further, the category document itself may be subject to categorisation.

This is exemplified in Figure 1: The article of the lemma '*South America*' is categorised by a category document of the same name. This document contains a list of articles it categorises which among others include the article we came from. The category '*South America*' is, in turn, hyponym of the hyperonyms '*Americas*', '*Continents*' and '*Latin America*'. These categories belong to a chain of hyponym relations which eventually lead a root category '*Categories*'.

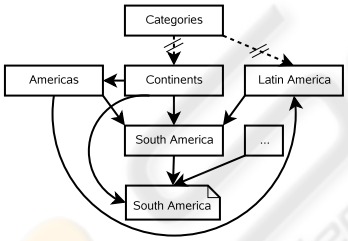

Figure 1: A sample of the wiki category system related to 'South America'.

2.2 Graph Representation

The small sample of Wikipedia in Figure 1 already suggests that the category system (i.e. the set of category documents and their interrelations) is more complex than a simple tree-structure. In order to grasp this complexity and its peculiarities we map the category structure onto a directed graph. The mapping is straightforward: We regard category documents as vertices and their hyperlinks – which constitute, for example, hyperonym relations ('*hyperonymOf*'), as directed edges (*from* a putative hyperonym *to* a corresponding hyponym).

Table 1 shows some general statistics of the German distribution of the Wikipedia. On 2006-08-03 it contained 415,980 articles of which about 94.2% were assigned to at least one of the 30,690 categories of that release. Figure 2 shows the distribution of assigned categories per article. The mean value is 2.57 categories per article whereas the standard deviation is 2.04 indicating that the distribution is quite stable.

The fact that over a million categorisations where

Table 1: Characteristics of the German Wikipedia.

Nodes Total	446,670
Articles Total	415'980
Categorised Articles	391,837
Uncategorised Articles	24,143
Categories	30,690
HyperonymOf-Relations	43,078
Categorises-Relations	1,069,005
Root-Categories	7,028
Number of cycles	16

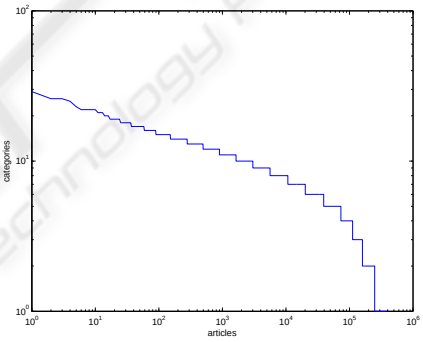

Figure 2: Distribution of assigned categories per article (log-log scale).

made shows how well this mechanism is accepted and applied by the community. This does not say anything about quality whatsoever. In fact there are a few characteristics which hint that the category system is a bit messy. First of all one would expect that a well formed category system has *one* designated root and an acyclic graph structure. Most Wikipedia distributions, including the English and German one have a designated root called 'Categories' or 'Hauptkategorie', respectively. However, there are $7,027$ categories in addition which do not have a parent hyperonym and thus form alternative roots. In terms of graph theory and supposing that each of these roots would dominate a single tree, we would need to speak of a forest. But as we actually deal with graph-like structures which consist of kernel rooted trees, we face the situation of a forest of generalised trees instead (see below). Moreover, an analysis of all paths between the category documents revealed that there are 16 *cycles*.

Thus, the wiki category system does, clearly, not instantiate a hierarchical taxonomy.

To complete the picture, we examined the distribution of in- and out degrees of the category nodes (cf. Figure 3): The mean in-degree is 1.4 and the standard deviation equals 1.03. The distribution of the out-degrees shows a mean of 36.24 hyponyms per category and a standard deviation of 568.77, thus pointing to a high variability as expected by scale-free phenomena. Why does the category system of Wikipedia

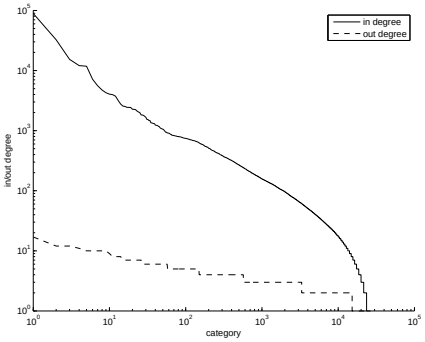

Figure 3: Distribution of category in/out-degrees (log-log scale).

make a rather chaotic impression – *far from the order of a hierarchical taxonomy*? The main reason might be given by the collaborative approach. Any user can specify any categorisation and hyponym-/hyperonym relation as he or she wishes. This has two important implications: On the level of article categorisations, there does not exist a common agreement what categories apply and how strict a definition (if one informally exists) has to be taken. This leads to extreme cases where an article (namely the one about Johann Wolfgang Goethe) is associated to 29 categories – in this sense polysemous categorisation is a phenomenon to be observed naturally in social ontologies. Another implication relates to the 'inner' organisation of the category system. We assume that only few if any users have a complete overview of the categories and how they are associated. Especially users which are new to Wikipedia might tend to build up some idiosyncratic category sub-structure for the domain they intend to write about – even though a similar one may already exist. One way to tackle this problem would be a semi-automatic supervision – which is a hopeless task given the dynamics in which Wikipedia is being edited. Since we cannot expect this situation to change in the near future, we have to think about representation models which enable categorical views without disregarding the factual complexity of the wiki category system.

2.3 From Raw Data to Generalised Trees

Our goal is to use the category system of Wikipedia to support the construction of corpora of a specific domain. To roughly sketch a scenario consider the task to build a corpus for a classification experiment which has to distinguish documents which belong to different subcategories of a common domain. The overall category might be 'Sports' whereas the subcategories (of which the articles have to be separated) belong to 'Motor Sport', 'Team Sports' and 'Individual Disciplines'. To transfer this task to Wikipedia you could try to pick the respective category 'Sports', check if it has equivalent subcategories and select those articles which belong to them. This and related approaches demand a tree-like structure of the category system which, as we have demonstrated above is by far not the case: The structure contains several roots as well as cycles and quite a number of categories have more than one hyperonym. *How, then, can a tree be extracted?*

A first approach would be to build a spanning tree of the category graph, that is all edges are removed from the graph until the constraints of a tree (or a forest) are met. The problem of this solution is that there typically are numerous trees which can be derived from the basic graph which are equally valid. Figure 4 illustrates this situation. Let G be the graph representation of an exemplary category system. Vertex 1 is the only vertex which does not have an incoming edge which makes it the only candidate for the root. Starting from this vertex, at least three different trees can be derived by discarding one or more edges which – due to the loss of information – result in different semantics.

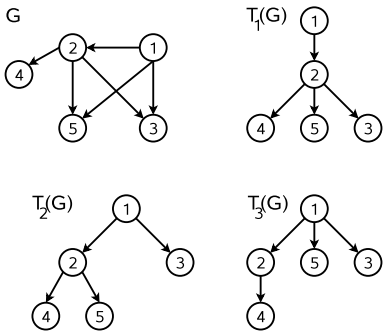

Figure 4: Variants of spanning trees for a graph G.

Since a loss of semantic information is unavoidable in the case that a specific tree-structure is selected over its equally selectable alternatives, a representation model is needed which overcomes this risk by incorporating the complete underlying edge set.

That way one may quickly choose a different heuristic to construct a tree based on the underlying category graph. In order to do this we adopt the notion of a *generalised tree*. The starting point of this approach is to utilise an edge typing starting from a graph's kernel hierarchical skeleton which is said to be spanned by so called **kernel** edges (cf. example in Figure 5). More specifically, we type edges as kernel which constitute the tree structure that was extracted by a specific spanning tree algorithm. The typing of the remaining edges is based on this initial step:

- **down** links associate nodes of the kernel hierarchy with one of their (im-)mediate successor nodes in terms of the kernel hierarchy.
- **up** links associate analogously nodes of the kernel hierarchy with one of their (im-)mediate predecessor nodes in terms of the kernel hierarchy.
- **across** links associate nodes of the kernel hierarchy none of which is an (im-)mediate predecessor of the other in terms of the kernel hierarchy.

For an in-depth description cf. (Mehler and Gleim, 2006). By representing the wiki category system as a generalised tree, none of the category-links is disregarded, but made accessible for further processing when it comes to extracting corpora of similarly categorised wiki articles.

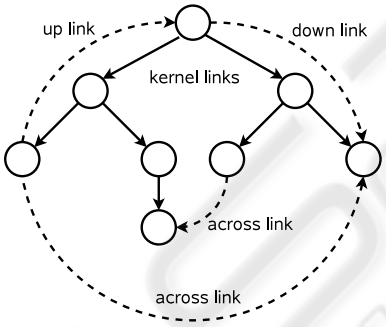

Figure 5: A sample generalised tree.

3 CATEGORY-BASED ARTICLE SELECTION

The previous Section has introduced the concept of kernel hierarchical structures and how it can be used to tackle the flaws of the category system of Wikipedia by representing it as a generalised tree. The advantages of the representation are that on the one hand an easy to process kernel-tree is established while at the same time all hyperonym relations are still available. Based on these preliminary steps several applications are possible. In this section we exemplify how the kernel hierarchical structure of a category system can be used to create categorical views and how category-based article subsets of Wikipedia may be selected. Section 4 will describe an implementation of the concepts in form of a browsing and extraction tool. Further applications which benefit from these results are discussed in Section 5.

3.1 Categorical Structure

The motivation of this section is to enable an information scientist to gather a subset of Wikipedia articles in order to build domain specific corpora. We start our considerations with the simple task to select articles which belong to a category that marks the boundary of the domain, say 'Jazz'. Figure 6 shows an excerpt of the kernel tree including associated articles. Note that the categories shown in the example may have much more interrelations which are not kernel (e.g. up links or across links) and not displayed. In case of the German distribution of Wikipedia there are 86 articles which are directly categorised as 'Jazz'. For some studies this may be enough, but often – especially in quantitative linguistics – more instances are desirable. Therefore, we extend the subset by also including those articles which are mediately categorised as Jazz (e.g. via subcategory 'Jazz-Style'). Based on the kernel hierarchical structure of the Wikipedia category system this is done as follows: Let T_J be the kernel-subtree of which the node representing category 'Jazz' is root. Iterate over all nodes t_i of tree T_J and add all articles which are categorised by category t_i to the result set.

The approach to select all articles which are immediately or only mediately categorised by a given category is quite exhaustive. On the other hand the set of selected articles might well be too heterogenous and may range from articles about different sub-genre of Jazz over musicians to festivals and typical instruments. In order to restrict the range of selected articles to a more specific subdomain one may go deeper into the category tree, that is consider child nodes of the category node that has initially been in focus. If that is still too general one may go a step further and so on. This concept of category-based article selection may be parameterised furthermore by distinguishing the multiplicity by which the considered articles are categorised. In general articles may be assigned to an arbitrary number of categories. However, one might restrict the selection of articles to those who are uniquely categorised or at least with respect to those categories which are (im-)mediate hyponyms of the selected root category.

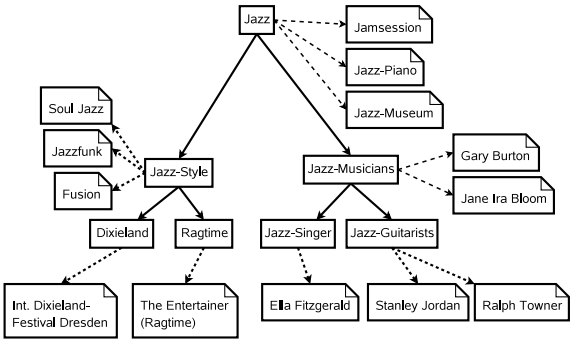

Figure 6: Excerpt of category graph including associated articles.

3.2 Article Characteristics

The approach of article selection presented so far has concentrated on the structure of the category system and how focal articles are located within. As far as selected articles are subject to machine learning studies it is desirable to filter out those articles which do not meet certain quality factors. The notion of 'article quality' invites a wide range of different interpretations. Instead of arguing about the impact of collaborative approaches on the quality of text production we take a pragmatic view by selecting articles which obey some "norm" with respect to their size, lifespan, number of revisions and related characteristics.

Revisions and time span The number of revisions of an article gives a first idea how controversial its content is. More specifically, we can ask for the time span between the first and last revision of an article and compute several expected values with respect to this time span as well as with respect to the number of revisions. Combining these criteria help to indicate how well an article is 'settled' into the Wikipedia.

Distinct contributors One might argue that an article which has been online for a long period of time and was edited more than once may nevertheless be of poor quality. In the case of so called 'edit wars', where few authors undo each others contributions, a large number of revisions is produced which in the end may be prejudicial to article quality. There may also be cases where articles have been edited several times but always by the same author. This is another example where solely relying on the number of revisions and the lifespan may fail. Therefore, we also take the number of distinct contributors per article into account and compute its expected value and standard deviation in order to rate the corresponding observed value (see below).

Content size A final characteristic we take into account is article size. That way articles are filtered out which barely contain any content or, inversely, are too long by integrating irrelevant material with respect to the definition of the focal lemma. From the point of view of machine learning, it is the former case which poses the problem of insufficient positive information, while the latter relates to the problem of supplying irrelevant or even false information.

We do not expect the number of revisions or distinct contributors to grow proportional to article quality. Instead, we compute the mean and standard deviation of each characteristic in order to span a *window* which filters out articles of 'unusual' characteristics. That is, for a given variable X (e.g. size, lifespan etc.) we compute the mean μ_X and standard deviation σ_X and get the interval

$$\mu_X - \frac{\sigma_X}{\nu} < x < \mu_X + \frac{\sigma_X}{\nu}$$

in order to select those articles whose observed value x of X falls into this interval ($\nu \in [1,\infty]$ is a parameter of our approach). Next, we define a selection rule $R_X : A \to \{0,1\}$ with $R_X(a) = 1 \iff \mu_X - \sigma_X/\nu < X(a) < \mu_X + \sigma_X/\nu$, A is the set of all input articles and $X(a)$ is the observed value of variable X in article $a \in A$. Finally, we build a constraint schema $\mathcal{R} = \{R_{X_i} \mid i \in I\}$ which allows to select all articles a for which $R_{X_i}(a) = 1$, $i \in I$. That way we can select for a given set of categories all instance articles whose size, number of revisions and lifespan lie within a certain interval around the corresponding mean value.

4 THE WikiCEP

The previous sections have described an approach to utilise the Wikipedia category system in order to provide graph-like browsing facilities and category-based article selections. In this section we describe in depth the tool we have built in order to implement this functionality. We give a short overview of its preprocessing steps and demonstrate its key features. Since the algorithmic background has already been introduced we focus on its user interface.

The first question that rises when it comes to processing the category system of Wikipedia is: *How to get it?* In contrast to the built-in mechanism to browse through the categories in Wikipedia (i.e. the underlying Mediawiki-Software), our tool relies on an offline representation of its category system. The Wikipedia Foundation offers XML-dumps of all distributions which come in different flavours. The variants mainly differ in coverage. The largest one includes all

document types (e.g. articles, categories, talks, portals) and all revisions in full text- so it comes to no surprise that the *compressed* file of the German distribution is about 16GB in size. However, since the file also contains meta information (e.g. the names of the contributors, the time a revision took place etc.), it is worth the trouble. In any case, the interlinking of the documents – including the category-related links – are not explicitly stored but must be manually parsed from the document source codes which are embedded into the XML- document.

In order to preprocess the raw data and extract the information which is relevant to the link structure and meta data we haven written a separate tool. Its purpose is to read the XML-dump using an efficient SAX-Parser and parse the meta data and document sources, reconstruct the interlinking and store the information as an attributed, typed graph. We use the XML-based *Graph eXchange Language* (GXL) (Holt et al., 2006) which allows to represent hierarchical hypergraph structures of arbitrary complexity. The GXL representation contains all information needed as input to the WikiCEP. However, the file still is about 11GB in size (note that the contents are not stored but only the structure!) which is still too large to efficiently load it. Therefore, we create a compact representation based on the GXL-file which solely contains the relevant information for grasping the category system and the corresponding article categorisation.

The GUI of the WikiCEP is organised into two sections: The **Main Category** (cf. Figure 7) tree view allows to browse through the category system represented as a kernel hierarchical structure. Technically speaking, it is an extension of the Java Class JTree which allows not only to visualise trees but also *graphs*. If a node of this generalised tree view has any links to some categories or articles which are not kernel, they are represented by icons left to the documents name (separated by in- and out-going links). A tool-tip shows the list of non-kernel (i.e. across, down or up) links of each node. By clicking on the respective icon it is possible to jump to the destination hyponym or hyperonym. *This enables the user to freely navigate through the complete graph structure of the wiki category system.* A double click on one of the nodes opens the respective document online in a separate browser. Finally, a (substring) search function allows to quickly locate categories as well as articles.

If a category is selected in the main category tree view, the respective node becomes the root of a **category subtree** (cf. Figure 8) which can be used to select articles for extraction according to the criteria explained in Section 3.2.

By default, all immediate (kernel-)subcategories

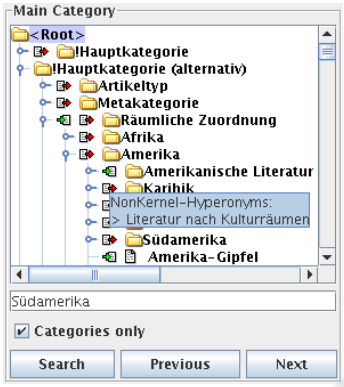

Figure 7: Generalised tree view of category system.

of the selected main category are marked for extraction. Performing an extraction now would result in the creation of a directory structure in the file system representing the main category and a set of child categories by analogy to the marked subcategories. Then all articles which directly *or indirectly* belong to the subcategories are extracted and stored accordingly in the directory structure.

The set of articles to be extracted can further be filtered out. First of all, it is possible to select whether the desired articles should be uniquely categorised, unique with respect to the subcategories of the main category or whether no restriction should be made at all. Furthermore, it is possible to select whether the articles to be extracted should directly be categorised by the subcategories or if a more lenient rule should apply (i.e. mediate categorisation). Finally, a slider allows to specify a window around a combined median of the statistical article features as described in Section 3.

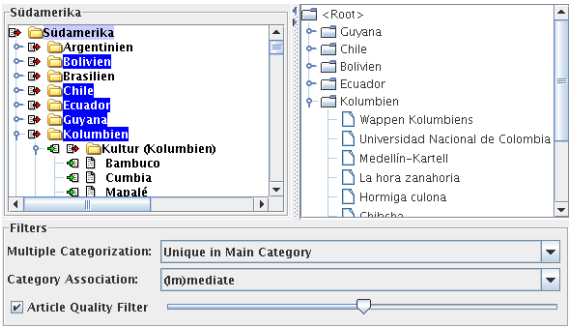

Figure 8: Category and article selection for extraction.

Currently, only one heuristic to compute the kernel hierarchical structure is implemented. However alternative approaches are in development.

5 APPLICATION SCENARIOS

So far we introduced an approach to representational and algorithmic issues of exploring wiki category systems. The implementation of the WikiCEP reflects these considerations. It supports researchers who need to gather corpora for their machine learning tasks. In this section, we outline three of them:

- *Text categorisation* is the task of automatically assigning category labels to a set of input texts (Sebastiani, 2002). It hinges on the availability of positive and negative training samples in order to train reliable classifiers. One way is to use the input corpus in order to separate training and test data and to overcome its limited size by means of cross-validation methods (Hastie et al., 2001). We propose using the WikiCEP as a means to additionally select data or to enlarge the feature space by exploring similarly categorised articles.
- *Lexical chaining* is the task of exploring chains of semantically related words in a text, that is, tracking semantically related tokens (Budanitsky and Hirst, 2006). It hinges on the availability of terminological ontologies like WordNet. We propose using the WikiCEP as a means to explore the Wikipedia category system as a social terminological ontology instead, that is, we propose using the Wikipedia as a source of defining semantic relatedness and similarity of lexical units.
- In *lexicology*, corpora are widely used for various applications. This relates, for example, to harvesting for new lexical terms, word sense disambiguation and the extraction of exemplary phrases. (Kilgarriff et al., 2005) describe the development of a corpus to support the creation of an English-Irish distionary which, besides print media, incorporates web documents. Further, (Baroni and Bernardini, 2004) propose an approach to incrementally build specialised corpora from the web based on a set of seed terms. WikiCEP marks a complementary approach which enables lexicographers to incorporate Wikipedia articles for their work.

6 CONCLUSION

This article addressed the potential of social tagging which Wikipedia offers to classify articles in order to enhance browsing for readers as well as to support the composition for domain-specific corpora. We mapped the category system onto a forest of generalised trees as an enhanced representation format for graph-like structured ontologies. This, nevertheless, allows tree-like processing of the data while keeping full information and overcoming flaws like cycles and multiple root categories (by introducing a virtual root to the kernel structure if necessary). Section 3 and 4 showed an exemplary application of the enhanced representation of the category system which adresses composition of domain-specific corpora and enhanced browsing. Future work will address the utilisation of more sophisticated heuristics to build the kernel hierarchical structure.

REFERENCES

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the LREC, Lisbon*.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning. Data Mining, Inference, and Prediction*. Springer, Berlin/New York.

Holt, R. C., Schürr, A., Elliott Sim, S., and Winter, A. (2006). GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, 60(2):149–170.

Kilgarriff, A., Rundell, M., and Dhonnchadha, E. U. (2005). Corpus creation for lexicography. In *Proceedings of the Asialex, Singapore, June*.

Leuf, B. and Cunningham, W. (2001). *The Wiki Way. Quick Collaboration on the Web*. Addison Wesley.

Mehler, A. (2006). Text linkage in the wiki medium – a comparative study. In *Proceedings of the EACL Workshop on New Text – Wikis and blogs and other dynamic text sources, Trento, Italy, April 3-7*.

Mehler, A. and Gleim, R. (2006). The net for the graphs – towards webgenre representation for corpus linguistic studies. In Baroni, M. and Bernardini, S., editors, *WaCky! Working Papers on the Web as Corpus*, pages 191–224. Gedit, Bologna.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Shapiro, A. (2002). Touchgraph wikibrowser. http://www.touchgraph.com/index.html.

Zlatic, V., Bozicevic, M., Stefancic, H., and Domazet, M. (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks. `http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:physics/0602149`.