Using community-generated contents as a substitute corpus for metadata generation

M. Meyer*

SAP AG, SAP Research CEC Darmstadt,
Bleichstr. 8, 64283 Darmstadt, Germany
E-mail: marek.meyer@sap.com
*Corresponding author

C. Rensing and R. Steinmetz

Multimedia Communications Lab,
Technical University Darmstadt,
Merckstr. 25, 64283 Darmstadt, Germany
E-mail: christoph.rensing@kom.tu-darmstadt.de
E-mail: ralf.steinmetz@kom.tu-darmstadt.de

Abstract: Metadata is crucial for reuse of Learning Resources. However, in the area of e-Learning, suitable training corpora for automatic classification methods are hardly available. This paper proposes the use of community-generated substitute corpora for classification methods. As an example for such a substitute corpus, the free online Encyclopaedia Wikipedia is used as a training corpus for domain-independent classification and keyword extraction of Learning Resources.

Keywords: e-learning; classification; metadata; metadata generation; Wikipedia; substitute corpus.

Reference to this paper should be made as follows: Meyer, M., Rensing, C. and Steinmetz, R. (2008) 'Using community-generated contents as a substitute corpus for metadata generation', *Int. J. Advanced Media and Communication*, Vol. 2, No. 1, pp.59–72.

Biographical notes: Marek Meyer works as a Researcher for SAP Research and for the Multimedia Communications lab at the Technical University Darmstadt. He has worked on concepts for improving reuse in e-learning, especially modularisation and aggregation of Learning Resources. Currently, he researches methods for automatic generation of metadata. He got his Diploma in Computer Science from the Technical University Darmstadt in 2005.

Christoph Rensing works at the Multimedia Communications lab at the Technical University Darmstadt. He leads the Knowledge Media group. His research interests are authoring support in e-learning, content repositories, repurposing of e-learning contents and web 2.0 technologies for learning. He has a Diploma in Information Management and a PhD in Computer Science.

Ralf Steinmetz worked for over nine years in industrial research and development of distributed multimedia systems and applications. He has been Head, since 1996, of the Multimedia Communications lab at the Technical University Darmstadt, Germany. From 1997 to 2001, he directed the Fraunhofer (former GMD) Integrated Publishing Systems Institute IPSI in

Darmstadt. In 1999, he founded the Hessian Telemedia Technology Competence Center (httc e.V.). His thematic focus in research and teaching is on multimedia communications with his vision of real 'seamless multimedia communications'. With over 200 refereed publications, he has become ICCC Governor in 1999, was awarded the ranking of Fellow of both, the IEEE in 1999 and the ACM in 2002.

1 Introduction

E-learning – enhancing learning by usage of computers – has become more and more successful over the last years. However, the production of high-quality digital materials, which are called Learning Resources throughout this paper, is still very expensive. Therefore, one of the key factors to the success of e-learning is re-use of Learning Resources, as multiple usage leads to more efficiency. If a teacher wants to re-use a Learning Resource, which has been produced by another person, he has to find and obtain it first. For this purpose, Learning Object Repositories (LORs) have been built. These LORs contain several Learning Resources of different authors. Each Learning Resource is described by a metadata record that contains the most relevant information about the Learning Resource. The better the metadata describes the Learning Resource, the better it can be found. Thus, metadata quality has a high impact on findability.

Unfortunately, creation of good metadata is often neglected by authors of Learning Resources. A manual creation and maintenance of a catalogue by librarians is not feasible because of the large number of Learning Resources in today's LORs. This is where metadata generation and metadata extraction enter the stage (Bergstraesser, 2005). Metadata can be created by several analysis methods that consider the contents of a Learning Resource. Especially the generation of topic-related metadata, such as keywords and categories, is important, because most users search for Learning Resources by topic.

Ideally, keywords are not only some words that are somehow related to the contents, but represent concepts, which are covered by a Learning Resource. This requires an ontology that contains all possible concepts. As Learning Resources in a repository are typically located in very different knowledge domains, a really comprehensive world ontology is needed. An ontology or taxonomy is also used for classification. In some repositories, classification is restricted to exactly one class per Learning Resource; other repositories allow multiple classifications.

This paper analyses existing approaches for domain-independent keyword extraction and classification in e-learning and proposes the use of community-generated substitute corpora as a new approach. The method described in this paper is based on the free encyclopaedia Wikipedia, which meanwhile contains more than 1 million articles on a multitude of topics. Wikipedia is probably the most complete, freely accessible and structured collection of world knowledge in the internet.

Three hypotheses are formulated in this paper as a foundation for the approach. Starting from the assumption that a Wikipedia article on a particular topic has a certain similarity to Learning Resources on the same topic, we introduce a practical implementation approach and present first results. The approach is based on standard machine-learning methods, but with a novel data source.

This paper is structured as follows: First some related work regarding metadata extraction is discussed in Section 2. Community-generated contents are addressed in Section 3. In Section 4, a new approach for topic detection based on Wikipedia is introduced. Section 5 gives an overview on the characteristics of the Wikipedia collection, discusses performance issues and presents a practical approach for a first implementation. Classification results of that implementation are presented and discussed in Section 6.

2 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 (Noufal, 2005), and for Learning Resources, in particular (Bergstraesser, 2005). 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.

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 analysing the contents of a document, whereas process metadata, such as author and creation date, can be obtained from the authoring environment (Hoermann et al., 2005). In the following, existing methods for content-related metadata extraction will be discussed.

Content-related metadata is the most important type of metadata for retrieval of documents and especially Learning Resources. Users search more often by words that describe the desired contents than, e.g., by a creation date or author name. Common content-related metadata fields are title, keywords, classification and an abstract or brief description. Using these fields is usually more efficient than using full text search and produces more relevant search results. The discussion of content-related methods will focus on keywords and classification. Keywords are terms that give a hint on the topics that are covered by a document; these keywords can be any words without restrictions. Classification, in contrast, is restricted to a fixed taxonomy or ontology, from which concepts can be taken to describe the contents of a document. Hence, the methods for generation of keywords and classification information also differ: for classification a mapping to known terms is required, whereas arbitrary words may be produced as keywords.

Classification problems are addressed by classification and clustering methods. Classification here means again that a document is assigned to one or multiple predefined classes. Clustering algorithms build new classes based on the similarity of documents.

Classification methods are a traditional focus of machine-learning technologies. If a large enough set of classified examples – also called corpus – is available, it can be used for training a system to automatically assign new documents to the existing classes. Examples for such systems are artificial neural networks, Support Vector Machines (SVMs) and the nearest neighbour algorithm. These methods are also called supervised learning, because desired outputs are known in the training phase. Another approach for classification of documents is rule-based systems, such as the ontology-based metadata generation described in Stuckenschmidt and van Harmelen (2001).

The unsupervised equivalent to classification is clustering. Clustering algorithms calculate a distance between documents and build groups of documents, which are near to each other or have common attributes. A common clustering method is the *K*-means algorithm. A method for clustering newspapers is presented in Newman et al. (2006). A set of 400 clusters is calculated based on the co-occurrence of entities, such as persons, organisations and places. Each of these clusters represents a hot topic that has been extensively discussed in the media.

Latent Semantic Indexing (LSI) is a probabilistic method, which is similar to clustering. The term-document matrix of a set of documents is transformed into a low-rank approximation by merging terms to concepts. This transformation can be used to calculate the covered concepts of a document. However, the concepts produced by LSI do not necessarily have a real meaning. Therefore, LSI is not suited for generating human-interpretable classes or keywords (Newman et al., 2006). Similar to LSI is the Random Indexing method, which lowers rank by using random dimensions (Sahlgren, 2005). Random Indexing provides comparable results to LSI, but avoids the complex calculation of term co-occurrences.

Some approaches for classification of documents have been presented above – methods for keyword extraction will follow. Keyword extraction methods can be classified by their coverage: Domain-dependent methods are limited to a particular knowledge domain but usually provide better results. Domain-independent keyword extraction methods can be applied universally, but are less precise. Domain-dependent methods are based on a domain model, which contains relevant terms for the particular domain. Documents are searched for these terms for determining keywords. Kruschwitz (2001) demonstrates how to build a domain model out of existing documents. Matsuo and Ishizuka have introduced a domain-independent method for extracting keywords from a single document without having a large corpus of documents (Matsuo and Ishizuka, 2004). This approach is based on the specific distributional characteristic of terms.

Another technology for extracting keywords from web pages is to exploit the structure of a document (Kruschwitz, 2001). Opposing the approaches above, Kruschwitz uses only those terms as keywords, which appear in at least two different contexts within a document; the considered contexts are meta information, document headings, document title and emphasised parts of a document.

To summarise this section, useful metadata generation methods exist for classification if a large corpus of documents is available for training. In the area of keyword extraction, some domain-independent approaches exist, but most of them also depend on a training corpus of exemplary documents. Only the method of Matsuo and Ishizuka works domain-independently on a single document. Some statistical methods can also be applied for keyword extraction in a domain-dependent case.

3 Community-generated contents

A lot of regional and global communities have emerged since the invention of the internet and the World Wide Web. In the early times of the internet, communities were rather small and passive. There were only few people using the internet, online time was expensive and participating by producing own contents, such as static HTML pages, was difficult. Users mainly consumed contents that were produced by professional content authors.

Over the years, the World Wide Web and its users have evolved in several aspects. The number of users has grown significantly; in many regions of the world, a majority of the population has access to the internet. Broadband connections are available at low costs. The creation of contents has also become much easier. First, there were local tools for creating HTML pages. Afterwards, new kinds of internet applications appeared that enabled users to enter contents without caring about details of HTML or other content formats. It started with guest books and forums and continued with consumer reviews about products, Wikis, Blogs and collaborative tagging. Passive consuming users have turned into active amateur authors. O'Reilly (2005) has coined the term 'Web 2.0' for these new applications.

Today, a large number of online communities exist that create their own contents without being professional authors or having a commercial interest. Such communities need a critical mass to create and maintain the contents. The size of the critical mass depends on a community's goals. The larger a community becomes, the larger the contents may grow. Most communities are open: anybody may participate by contributing and using the contents.

One successful community project is Wikipedia (2006a). The goal of the Wikipedia community is to create a free online encyclopaedia. It is based on Wiki software. Anybody may contribute to the project by writing, updating and extending articles. The encyclopaedia is free to use for everybody. Wikipedia has meanwhile reached the size of commercial encyclopaedias and become one of the most frequently visited websites in the world.

4 Using Wikipedia as a substitute corpus

As the previous section has shown, classification of Learning Resources based on topics requires two prerequisites: predefined topic classes and a training corpus, which contains several examples per class. There are good exemplary corpora for newspapers for web pages, e.g., the Reuters corpora for news or the TREC corpora for web pages (Lewis et al., 2005; Commonwealth Scientific and Industrial Research Organization, 2006). For e-learning repositories, however, there is no suitable corpus yet. One major problem of current e-learning repositories is that they contain too few Learning Resources. Combined with the fact that Learning Resources are not restricted to a certain knowledge domain, but may cover any topic, there is only little hope that a suitable Learning Resource corpus for automated topic classification will be available in the near future.

Therefore, a new approach is proposed by this paper. Instead of a real corpus of Learning Resources, a substitute corpus shall be used, whose entities bear enough resemblance to Learning Resources. As described above, the lack of an extensive training corpus prevents effective classification of Learning Resources. But several community projects exist, which generate huge amounts of explicit knowledge. If a data source can be found, whose entities are similar enough to Learning Resources regarding a set of relevant attributes, this data source could be used as a substitute corpus for classification methods.

The free encyclopaedia Wikipedia (2006a) is suggested as such a substitute corpus. Wikipedia is a free, web-based encyclopaedia, which is written and updated by a large community of volunteers. This large community ensures that all topics that seem relevant to anyone already are or probably will be described by a Wikipedia article. Wikipedia is also available in several languages, is continually updated and still grows over time. By April 2006, the English Wikipedia database contained more than one million articles. But Wikipedia is not just a collection of articles: it also provides a classification system: Each article may be assigned to one or more hierarchically organised categories.

The important research question is: Is Wikipedia suitable as a substitute corpus for Learning Resources? This paper addresses this question and works towards an answer. The underlying general hypothesis is:

> **Hypothesis 1** *(General Wikipedia Hypothesis)*: *Learning Resources and articles of the Wikipedia encyclopaedia both are knowledge transfer texts. As such, they bear a resemblance. If a Learning Resource and a Wikipedia article cover the same topic, a similarity between them can be measured.*

If this similarity between Learning Resources exists, it should be exploitable by Information Retrieval methods. Therefore, statistical similarity measurements (e.g., a distance function in a document vector space) are used as a basis to formulate a more specific hypothesis.

> **Hypothesis 2** *(Specific Wikipedia Hypothesis)*: *Whenever a Learning Resource is statistically similar to a particular Wikipedia article, there is also a similarity in the covered topics. If the statistical similarity exceeds a certain threshold, the Learning Resource covers the same or a closely related topic as the article.*

Hypothesis 2 raises some additional questions. First of all, which statistical methods are suitable to deduce topic similarity from statistical similarity? Second, which minimal threshold value assures a sufficient accurate classification? And finally, the choice of topics and a definition of topic matching have to be defined.

Furthermore, there are large Learning Resources that cover multiple topics. For classification of these Learning Resources, an additional definition of subtopics is helpful. For this purpose, each contiguous extract of a Learning Resource is regarded as a Learning Resource fragment.

> **Hypothesis 3** *(Fragment Hypothesis)*: *Hypothesis 2 also applies accordingly to Learning Resource fragments. Whenever a topic has been determined as topic of a Learning Resource fragment, it is also considered to be a subtopic of the embracing Learning Resource.*

If these hypotheses are true, they can serve as a foundation of using the Wikipedia as a substitute corpus for Learning Resource classification. Two basic classification approaches based on Wikipedia articles are thinkable: Coarse classification using Wikipedia categories as classes or fine-grained classification by regarding each article – and thereby each individual topic – as one class. The second approach, regarding each article as a class, takes into account that each article addresses exactly one well-defined and disambiguated topic; the article title is suited for naming the class.

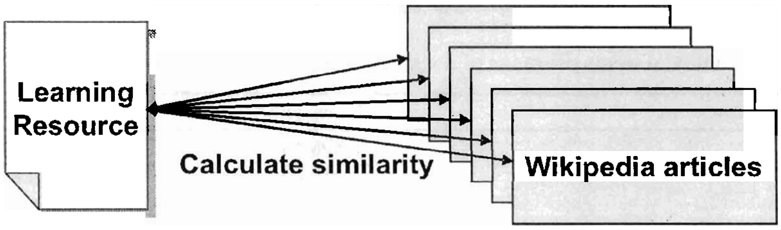
Keyword generation is a further application of Wikipedia-based topic determination. Keywords for a Learning Resource should be a very brief description of the contents. If matching topics are determined for all relevant fragments of a Learning Resource, the resulting topics and subtopics are very well suited as keywords. In contrast to most other methods, this approach does not depend on useful headlines or particular structures.

5 Proof-of-concept implementation

For testing the hypotheses, a first implementation of the approach has been realised. Main goal of the implementation is to prove Hypothesis 1 and identify critical factors for proving Hypotheses 2 and 3. Hypothesis 1 expresses that having a Learning Resource and a Wikipedia article about the same topic implicates a similarity of the two documents. That means that there is a co-occurrence of topic similarity and document similarity. The hypothesis though does not specify the way similarity is measured. In Information Retrieval, a common method for determining the similarity of texts is to compare them in a Vector Space Model (VSM) (Salton et al., 1975). Hence, this method is also applied for the first proof-of-concept implementation. For proving the first hypothesis, it is not necessary to identify the method that calculates similarity best; a method that provides a good similarity function is sufficient.

For a first test, a Learning Resource on the topic 'Network Calculus' is used. There is also a Wikipedia article available on that topic. To determine if there is a significant similarity between both documents, the Learning Resource is compared with all existing Wikipedia articles (cf. Figure 1). Hypothesis 1 can be assumed true if the similarity of the Learning Resource to the 'Network Calculus' article is significantly higher than the average similarity values. For proving Hypothesis 2, the 'Network Calculus' article would be required to be the best-matching article.

Figure 1 Basic approach for finding similar Wikipedia articles



As the Wikipedia-based topic detection approach is targeted not only for background tasks in repositories, but also for interactive classification and metadata generation applications, the implementation should be designed with regard to the run-time performance on workstations.

5.1 Analysis of Wikipedia

Before starting the design process, some characteristics of the Wikipedia encyclopaedia have been analysed. Up-to-date statistics on the size and usage of Wikipedia in different languages can be found online (Wikipedia, 2006b). This paper focuses mainly on the English version, and as comparison also on the German one. As of June 2006, the English Wikipedia contains about 1,300,000 articles; for the German Wikipedia version, 435,000 are listed. Most of these articles contain at least 200 visible characters.

For further analysis and an implementation, the provided complete database dumps have been downloaded. For the English version, the database dump from 20 April, 2006 is used. The German database dump dates from the 4 June. Once the database dumps are unzipped into plain XML files, they consume several Gigabytes of disk space (see Table 1). The dumps contain pages, which do not represent articles, but special pages, images pages or redirects – these pages are not regarded as articles throughout this paper.

Table 1 Statistics on Wikipedia

	English	*German*
Number of articles	1.3 M	435 K
Articles >200 char.	1.2 M	422 K
Average article size (bytes)	3133	3498
Size of database dump (zipped)	1.2 GB	524 MB
Size of database dump (plain)	5.2 GB	2.0 GB

Important for processing text documents in the VSM are terms. We consider only one-word terms; common stemming algorithms are used to map words to their basic word stem. We have counted the document frequency for each occurring term. The document frequency indicates in how many different articles a term occurs. Table 2 presents the number of terms that fall into different ranges of document frequencies for the English (EN) and German (DE) version of Wikipedia. In sum, there are over 3 million different terms in the English Wikipedia. But more than half of the terms occur in only one article. On closer examination, most of these terms seem to be words from different languages, fantasy words or unfamiliar names. All terms that occur in at most two documents form two-thirds of the whole vocabulary.

Table 2 Number of terms per range of document frequencies

Document frequencies	*Terms (EN)*	*Terms (DE)*
1	1,898,542	1,598,058
2	518,047	360,442
3	223,730	163,643
4–5	204,943	160,157
6–10	186,834	156,739
11–20	116,379	102,823
21–30	45,387	40,750
31–50	40,761	37,304
51–100	34,540	32,356
101–1000	43,495	39,514
1001–10,000	9,668	7085
10,001–100,000	2,421	1041
100,001–200,000	140	47
200,001–300,000	20	16
300,001–400,000	5	80
400,001–∞	7	2
Σ	3,324,919	2,699,985

5.2 Performance considerations

The implementation is considered to run on small servers and on typical workstations. We, therefore, assume a computer with a 3 GHz desktop CPU, 2 GB main memory and Java as programming language as the target platform. Owing to the operating system, overhead and other influences, only 1–1.5 GB of RAM are effectively available for an application. Furthermore, interactive metadata generation methods imply that a user is sitting in front of the computer and waiting for a metadata proposal to accept or reject; this leads to a desire for fast execution. Real-time behaviour – delivering results within some seconds – should be aimed at as optimum. Based on these conditions, performance considerations are discussed in this section.

The three most important reasons for performance bottlenecks are:

- disk memory
- size of in-memory representation
- structure of in-memory representation.

High consumption of disk space also causes many disk operations, which are very slow. The size of RAM footprint mainly matters if the amount of required memory exceeds the RAM size – in this case expensive swapping is needed. And finally, the structure of the run-time representation of data has an impact on the complexity of the comparison algorithm.

Assume that a non-optimised VSM implementation is used. The original English Wikipedia database dump is 5 GB. A rule of thumb for the dimension of a document classification indices is to multiply the size by 2; this leads to an index size of 10 GB – definitely too large for the target main memory. A memory size optimisation is required.

Second, consider the calculation effort to compute a simple distance function, e.g., the inner product of two vectors, in the VSM. For the ease of estimation, the number of articles is rounded down to 1,000,000 articles and the number of dimensions down to 3,000,000. A total number of 1000 Learning Resource fragments is assumed. As a result, 3×10^{15} floating-point multiplications have to be calculated. If one multiplication is executed per CPU cycle, the algorithm runs for about 277 h. This execution time would be inappropriate and has to be decreased.

Of course, some information retrieval libraries provide generic performance optimisations. But the known characteristics of the particular application can be utilised for a tailored optimisation. This means especially to find an accurate trade-off between memory consumption, execution time and classification quality. Objectives for the implementation are

- shrink run-time representation to fit completely into main memory
- optimise in-memory structure and complexity of algorithms for fast execution
- reduce consumption of disk space to minimise disk operations.

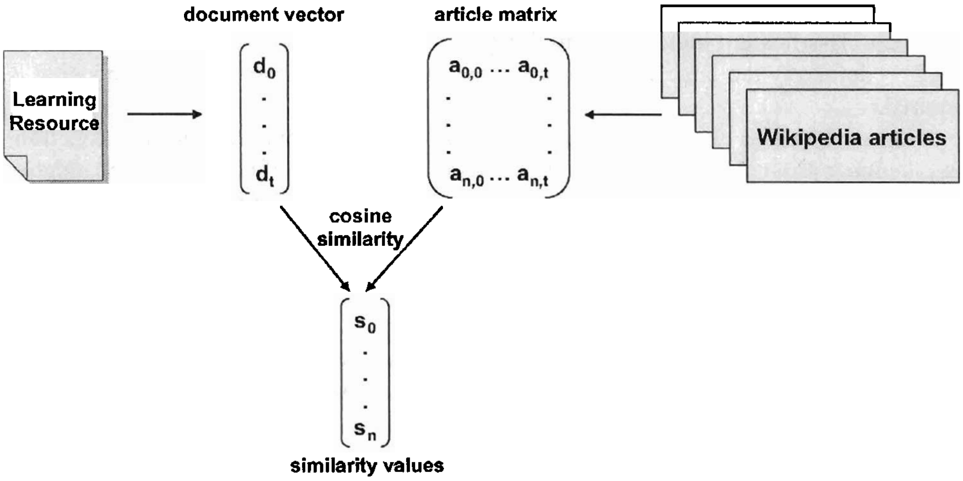
5.3 Implementation

The implementation is based on a VSM. All words that are used in any Wikipedia article are used as dimensions after a stemming algorithm and a stop list has been applied. Java 1.5 has been chosen as programming language.

In a preparation phase, all Wikipedia articles are transformed into document vectors. First, the whole Wikipedia database is scanned for all used words. The stemmed forms of these words are stored as a global word list that serves as a description of the vector space dimensions. In a second pass over the database dump, a document vector is created for each article. The vector is weighted by TF.IDF and normalised to a length of 1. The resulting article vectors are stored to disk. This preparation step has to be performed only once per version of the database dump.

At run-time, all article vectors are read into main memory again. Then, a Learning Resource is divided into several fragments. For each of these fragments, the contained text is extracted and transformed into a document vector; this transformation is performed analogous to the previous transformation of articles. Then the similarity values are calculated. Each fragment vector is compared with all article vectors. The cosine function (inner product) has been chosen as similarity function (see Figure 2). The highest matches for each fragment are interpreted as classification result. Depending on the mode, either a fixed number of matches or all matches with a similarity value above a certain threshold are used.

Figure 2 Calculation of similarity values



Top priority is to reduce memory consumption. First of all, a sparse vector representation is used, which means that only non-zero elements of vectors are stored. Considering that most articles contain only some 100 different words out of the vocabulary of 3 million words, the effect is significant. The dimension of the vector space may also be reduced to realise additional savings. This can be achieved by removing very frequent or rare words from the dimensions. A common method is to remove words with a very high document frequency, because they are considered to have only little relevance. Furthermore, Section 5.1 has shown that a very large amount of words occur in only one or two articles. On the one hand, rare words generally are considered to have a very high significance. On the other hand, in the case of Wikipedia articles, most of these terms seem to be names, words from a foreign language or misspellings. Also, a term, which occurs in only one or two articles, might be a rather unknown word. Removing terms with a document frequency of only one and two could significantly decrease the number of dimensions.

The run-time representation of article vectors is realised using the Compressed Sparse Row (CSR) format, which contains only non-zero values plus two index vectors: a column index, which determines the position within a vector, and a row index, which indicates where each vector starts (Goharian et al., 2003). An additional hash table for fast random access to non-zero values has been introduced. Based on the CSR representation for article vectors and a hash table representation of Learning Resource fragment vectors, an optimised algorithm for calculating the similarity with a lower complexity has been implemented.

6 Test results and interpretation

For the English and German Wikipedia, the database dumps have been transformed into a VSM for first tests. The used word lists have been varied to find out the effect of reduced vector space dimensions. For example, from the German Wikipedia all words that occur in less than three documents or more than 200,000 documents have been removed. The result was that the word list itself significantly shrinked, but the vector information decreased only slightly from 630 MB to 558 MB. Some sizes of VSM representations are given in Table 3. However, determining the impact on classification performance would require a large-scale experiment. For the English version, only a limited word list has been used because of limited main memory resources. The transformation process is very time consuming and took about two days for the German Wikipedia and four days for the English Wikipedia on a standard workstation.

Table 3 Size of Wikipedia articles after transformation into Vector Space Model

Language	*Used terms (doc. freq.)*	*Size of VSM data*
English	3–2 M	1.12 GB
German	All	630 MB
German	3–200 K	558 MB

In contrast, the run-time performance of the classification algorithm is much faster. Performing a classification of an English sample Learning Resource took about 30 min. However, the bottleneck is the transfer of data from disk into main memory. Twenty eight minutes were consumed by loading the article vectors, but only 24 s were needed for determining the similarity between a given Learning Resource vector and all articles. Creating a vector representation of a text document has taken less than 1 s. The total time of the method is quite high. But once the vector data is available in main memory, the classification works at an acceptable speed. If the article vector representations are persistently kept in main memory, the classification of a Learning Resource takes less than half a minute.

For the sample Learning Resource on 'Network Calculus', one vector for the whole course has been created. This vector has been used as input for the classification method. For testing Hypothesis 1, the similarity value for the article 'Network Calculus' was of interest. The measurement has produced a similarity of 39.66% for that article, whereas the average similarity value was only 1.17%. The most similar articles are listed in Table 4. This result supports Hypothesis 1. The category information of Wikipedia indicates the classes 'Network performance' and 'Computer network analysis', which can be followed upwards to the classes 'Network management' and 'Computer networking'.

Table 4 Result of topic classification for 'Network Calculus' course including category pages

Article	*Similarity*
Category: algebraic curves	0.445
Network calculus	0.397
Category: economics curves	0.320
Category: elliptic curves	0.285
Singularity (mathematics)	0.279
Singular points	0.278
Curved bar	0.272
Category: spirals	0.271
Category: packets	0.258

The classification results show some interesting characteristics. First of all, category pages – which had not been removed from the articles database before – are obviously overrated. This can be explained by the different linguistic structure of category pages: a category page consists mostly of titles of several related articles; therefore, it contains many high-rated keywords, in contrast to those words in natural language, which have less significance. Nevertheless, the classified categories are not completely off-topic, because the network calculus course uses mathematical curves for modelling network traffic.

If category pages are removed from the result list (see Table 5), 'Network Calculus' is the article with the highest similarity to the given Learning Resource. Furthermore, the next regular article in the list has a similarity value of only 28%, which is more than 10% lower than the correct match. If the top match is used, the Wikipedia-based classification method provides a good result for this sample course. Using a threshold of one-third (33%) shows the same result. For making a general statement on how to best select topics for metadata, another test with a larger set of Learning Resources has to be carried out.

Table 5 Result of topic classification for 'Network Calculus' course without category pages

Article	*Similarity*
Network calculus	0.397
Singularity (mathematics)	0.279
Singular points	0.278
Curved bar	0.272
Average similarity	0.017

7 Conclusions and outlook

This paper has introduced a new approach for metadata generation based on using the Wikipedia encyclopaedia as a substitute corpus. This approach uses standard Information Retrieval technology, but with a different data source. The test results that have been presented are very promising. Three hypotheses have been postulated as a foundation of the Wikipedia-based approach. The first two hypotheses have been supported by test results: The comparison of a Learning Resource with Wikipedia articles can be used for determining the topic of the Learning Resource. This information can be used either for classification or for generating keywords for a metadata record. For finally proving the hypotheses, further expanded tests will be required, including a consideration of statistical significance.

However, the approach might produce good results only for a particular granularity of Learning Resources. This has to be evaluated in further experiments. For a large Learning Resource, which covers a variety of topics, the approach might possibly produce less accurate results. Therefore, the third hypothesis becomes relevant. Subtopics for several fragments of a Learning Resource can be determined; then an overall topic is built out of fragment subtopics. The links between Wikipedia articles might be used to find the overall topic. Also, subtopics of fragments may be used as keywords for the Learning Resource metadata.

Some additional ideas have to be tested if they can improve the classification method. For example, different similarity functions, such as binary comparison or a word recall rate, have to be evaluated. Furthermore, if a topic domain of Learning Resources to classify is already known in advance, corpus for comparison could be limited to articles of that particular domain. A domain-limited corpus for classification of Learning Resources on medical science could for instance be much smaller than the general-purpose corpus.

The achieved results encourage one to watch out for other substitute corpora as well. In the course of the web 2.0 hype, a large number of community projects have been initiated for jointly creating immense amounts of freely accessible knowledge. Some of these projects have already reached a critical mass of contributors or will do so in the future. It seems promising to analyse these data sources for their qualification as input for alternative substitute corpora for classification tasks.

Acknowledgements

This work is supported by the German Federal Ministry of Economics and Technology in the context of the project Content Sharing.

References

Bergstraesser, S. (2005) *Automatisierung der Erstellung von Metadaten*, Diploma thesis, Darmstadt University of Technology, Darmstadt.

Commonwealth Scientific and Industrial Research Organisation (2006) *TREC-2004 Web Research Collections*, Online, last visited 28th August, http://es.csiro.au/TRECWeb/.

Goharian, N., Jain, A. and Sun, Q. (2003) 'Comparative analysis of sparse matrix algorithms for information retrieval', *Journal of Systemics, Cybernetics and Informatics*, Vol. 1, No. 1, pp.38–46.

Hoermann, S., Hildebrandt, T., Rensing, C. and Steinmetz, R. (2005) 'ResourceCenter – a digital learning object repository with an integrated authoring tool set', in Kommers, P. and Richards, G. (Eds.): *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2005*, AACE, Montreal, Canada, June, pp.3453–3460.

Kruschwitz, U. (2001) 'A rapidly acquired domain model derived from markup structure', *Proceedings of the ESSLLI01 Workshop on Semantic Knowledge Acquisition and Categorisation*, 17 October, Helsinki.

Kruschwitz, U. (2001) 'Exploiting structure for intelligent web search', *HICSS 2001: Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Vol. 4, IEEE Computer Society*, Washington DC, USA, p.4010.

Lewis, D.D., Yang, Y., Rose, T.G. and Li, F. (2004) 'RCV1: a new benchmark collection for text categorization research', *Journal of Machine Learning Research*, Vol. 5, pp.361–397.

Matsuo, Y. and Ishizuka, M. (2004) 'Keyword extraction from a single document using word co-occurrence statistical information', *International Journal on Artificial Intelligence Tools*, Vol. 13, No. 1, pp.157–169.

Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M. (2006) 'Analyzing entities and topics in news papers using statistical topic models', in Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B.M. and Wang, F-Y. (Eds.): *Proceedings of the IEEE International Conference on Intelligence and Security Informatics, ISI 2006*, Vol. 3975 of *Lecture Notes in Computer Science*, Springer, pp.93–104.

Noufal, P. (2005) 'Metadata: automatic generation and extraction', *7th MANLIBNET Annual National Convention on Digital Libraries in Knowledge Management: Opportunities for Management Libraries*, Indian Institute of Management, Kozhikode, pp.319–327.

Oreilly, T. (2005) *What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*, OReilly Media, Inc., http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Sahlgren, M. (2005) 'An introduction to random indexing', in Witschel, H.F. (Ed.): *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, August 16, Volume 87 of *TermNet News: Newsletter of International Cooperation in Terminology*, Copenhagen, Denmark, pp.1–9.

Salton, G., Wong, A. and Yang, C.S. (1975) 'A vector space model for automatic indexing', *Communications of the ACM*, Vol. 18, No. 11, pp.613–620.

Stuckenschmidt, H. and van Harmelen, F. (2001) 'Ontology-based metadata generation from semi-structured information', *K-CAP '01: Proceedings of the 1st International Conference on Knowledge Capture*, ACM Press, New York, NY, USA, pp.163–170.

Wikipedia (2006a) *The Free Encyclopedia*, last visited 28th August, Online, http://en.wikipedia.org.

Wikipedia (2006b) *Wikipedia Statistics*, last visited 4th September, Online, http://stats.wikimedia.org/EN/TablesRecentTrends.htm.