EMIS Collections and Conference Proceedings
FIZ Karlsruhe

INDEXING OF DOCUMENTS AND ITS IMPACT ON THE CHANGING ROLE OF MATHEMATICS INFORMATION SERVICES IN THE DIGITAL AGE

WOLFRAM SPERBER AND BERND WEGNER

Abstract. With the development of digital publications, mathematicians are confronted with lot of new challenges. Publishing mathematics is easier by applying TeX or other encoding systems. Access is possible wherever Internet is available. The mathematical literature is growing rapidly. Reference databases and search engines like Google help the users to find the publications in their fields of interest. In this context, an improved content analysis of mathematical information plays a key role. Standards, models, and algorithms are the most important ingredients for an efficient processing of information and retrieval. In this paper, some approaches are presented to improve and advance existing information services in mathematics by automatic tools for content analysis and a first prototype for a coarse automated key phrase extraction and classification is described.

1. Introduction

The availability of digital versions of publications in the Web has changed the framework for information sciences and services drastically. This concerns the basic concepts for the delivery of information as well as related services and tools. In more detail: All Information can be stored in digital form. It can be downloaded to computers, made accessible world wide, and linked with each other. Today, the Web can be considered as the largest memory in the history of mankind. Digital formats allow sophisticated presentations of information and processing of information by machines. The information in the Web is globally accessible. Information retrieval and access facilities have radically changed. Search engines are today the most important instruments for information retrieval. For this purpose, indexing, storage, and ranking of all (public) available information in the

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

Web like home pages or a publications has to be prepared. Text analysis and citation analysis are important tools in this context. On this background established printed documentation services like Zentralblatt MATH or Mathematical Reviews changed into reference databases. Providing subject-specific high-quality comprehensive information still play an important role today.

Mathematical Information Services looking back To describe the changes and challenges of mathematical information services, we start with a retrospective. The first comprehensive mathematical documentation service was founded at the end of the 19th century. Before some first catalogues indexing mathematical books were published, see, e.g., Scheibel [1] or Murhard [2]. The Crelle Journal [3] (today the 'Journal fr die reine und angewandte Mathematik') published reviews of mathematical publications starting with its first issue beginning in 1826. The foundation of the 'Jahrbuch ber die Fortschritte der Mathematik' (JFM) in 1868 [4] marked a milestone in the history of information services in mathematics. The first volume covered the bibliographic data of 875 publications, more or less the worldwide production in mathematics of the year 1868. This compares to a current annual production of more than 120,000 relevant articles and books in mathematics.

The JFM was a reaction of the mathematical community to objective needs like having an overview of the increasing size and bandwidth of mathematical knowledge:

- The number of mathematically relevant publications became too large. The JFM should help the mathematicians to identify and track the relevant developments in mathematics and build up a comprehensive memory of mathematical knowledge being identical with the printed mathematics publications at that time. As a first requirement, the bibliographic data of the publications had to be registered.
- This was accompanied by an enhanced content analysis: Reviews were added describing the content of the mathematical publications The Table of Contents (ToC) of the JFM was a first rudimentary instrument to assign information on its mathematical subject. Of course, the ToC was far from being some first version of a classification scheme in the formal sense, which did not exist at that time. The ToCs of JFM were not really stable, they changed from year to year. Hence one of the basic requirements for a classification scheme was violated.

The edition of the JFM was promoted by the mathematical community and supported by famous mathematicians like C. W. Borchardt, L. Kronecker, and K.

Weierstrass. Since that time, several additional approaches for editing a comprehensive information service in mathematics were undertaken like Valentin's catalogue (see [5]) and the dream of a comprehensive mathematical library (see [6]). But they failed to be successful for a longer period. The next break through was the foundation of 'Zentralblatt fr Mathematik und ihre Anwendungsgebiete' in 1931 with the aim to overcome some deficits of the JFM. The biggest one was the huge backlog of the JFM. Zentralblatt survived until nowadays as the database zbMATH [7]. In the following, we will use the notation Zentralblatt MATH when we talk about the editorial procedures or historical items and zbMATH when we talk about the current service.

Within its lifetime of more than 80 years Zentralblatt MATH experienced a lot of changes and developed several new concepts. Here we will concentrate on the aspects related to the content analysis of mathematical publications. One milestone was the implementation of and the indexing in accordance with classification schemes like the currently used Mathematics Subject Classification (MSC) [8]. This already started before the transition of Zentralblatt Math into the reference database zbMATH and lead to huge volumes publishing higher cumulated subject indexes. With the development of the database zbMATH, key phrases had been added from the text or by the subject editors with the goal to make them available as primary items for subject searches. Clearly all words appearing in a documentation entry were at the disposal for the so-called basic search. Up to now, the assignment of classification codes and key phrases is based on intellectual work done by the subject editors and the reviewers.

2. Current Challenges for Information Services

The digital presentation of mathematical publications has changed the framework for the production and the offer of mathematical information. There are new options to provide the information in a sophisticated way. This automatically confronts the providers with new needs and challenges:

- One topic are digital formats of mathematical information and the World-WideWeb (WWW). Most mathematical publications are accessible via the WWW now. The corresponding formats, steered by the W3C as HTML, XML, RDF, or OWL, provide standards for well-structured documents and allow enhanced options for representing and handling the information. Mathematical formulae are of high importance for mathematical publications. Markup formats like TeX allow the authors to create manuscripts, which are ready for publications. MathML allows

the processing of the information machines and the semantic enrichment of publications.

- Publishing not necessarily needs publishers anymore. Authors have the possibility to publish their manuscripts directly on the Web. We notice an increasing number of so-called grey publications. Some communities have established their own online archives. But in comparison with conventional publishing peer-reviewing is missing.
- A more than linear growth of the number of mathematical publications can be observed. At present about 120,000 mathematical articles and books are published annually. Clearly, this number depends where the borderline from mathematics to other sciences is drawn.
- A change of the retrieval behaviour has to be taken into account. Big search engines like Google, Yahoo, Bing, et. al. are becoming more and more popular among mathematicians. They are focused on general and comprehensive information, which can be generated automatically. Their aim cannot be to care systematically about complete and precise information on special topics, mathematics. This still will be the domain of reference databases.
- The types of mathematical knowledge available in the Web is going beyond mathematical publications, mathematical software being the most important example.

There are advanced user expectations for a mathematical information service, in particular from the mathematical community. Traditionally such a service should be as complete as possible, providing information on all relevant publications. This is still an important requirement. Also high-quality is expected by the users. The information should be validated and precise, the presentation should be well comprehensible, and the broader context of a document should be visible. It should be well integrated in the network of mathematical publications. Links on different levels should be available enabling access to the article itself, to those cited in the references of the article or even to other publications of interest in the same context.

To fulfil the last requirement was almost impossible, when no digital offers were available. Citations had to be looked up in libraries, where the corresponding publication was stored in print or on microfiche for example. With the availability of digital publications reference databases in mathematics could improve their service essentially. They were able to extend their role from a retrieval service to a navigation tool for digital mathematics libraries, providing an excellent facility

for their users to find information of interest in the network of mathematical publications. As a consequence, more time has to be spent for the preparation of the input for the reference databases in order to care about an enriched and extended content analysis and an enhancement of the data by an improved linking to relevant information. This only can be afforded if the work will be supported by automatic processing of the information. The following sections will describe how such a processing may look like for the semantic enrichment of the stored information.

3. Content analysis for zbMATH

We may distinguish between direct and indirect procedures for the content analysis. This paper will concentrate on the direct procedures. The indirect ones generally may be described as derivations from the captured data, possibly including the bibliographies, if they are available electronically. They include citation analysis, ranking, several kinds of profiles etc. A lot of efforts are spent at zbMATH for improving and running such procedures.

The main direct procedures are the preparation and capture of reviews or abstracts, the assignment of classification codes, the highlighting and capture of key phrases (including also mathematical symbols and formulae) and the standardization citations provided with the review or abstract. These procedures represent different levels of precision. Though being standardized, classification codes are a comparatively coarse concept for describing the subject of a publication. Key phrases may describe this in a finer way, but being uncontrolled they are of restricted importance for subject searches.

Reviews or abstracts should give a short impression of the content and the main achievements of a publication, relating them to other publications by top citations. Reviews are prepared on a voluntary basis by reviewers, i.e. invited experts from the mathematical community. They should provide an independent view of the paper on contrast to abstracts. Abstracts generally are written by the author or sometimes by an expert involved in the edition of the publications. They are controlled by the authors in any case. Top citations should refer to papers, which are strongly related to the one under consideration, going beyond the common citations in the bibliography.

The assignment of key phrases may be considered as a preparation of a very short summary of the paper by listing relevant mathematical terms. This procedure has been introduced for zbMATH rather soon after its first release as a searchable

database was available. It is still pure intellectual work by the subject editors, extracting phrases from the review or the abstract and adding relevant key phrases from other sources. These terms are not controlled, because no controlled vocabulary or thesaurus is available for mathematics. They hopefully may be used in the future to develop such a scheme.

Classifying publications in accordance with classification schemes is an old practice. This is a useful method to structure and identify relevant subjects in a special context. Classification schemes combine verbal descriptions of subjects with classification codes and have a tree-like structure, displaying different levels of generality. These different levels enable the user to apply the scheme to collections of different magnitude. They also comply with the requirements of the paper era where documents have to be totally ordered according to subjects and lexicographically by authors within the same subject.

There are several classification schemes for mathematics, having different levels of precision, depending on the purpose they had been developed for. Most of them are mathematical parts of more comprehensive classification schemes. They have the advantage that many relations of mathematics to other sciences could be represented more appropriately. The reference databases in mathematics currently are using the Mathematics Subject Classification MSC (zbMATH, MSN, CMA) and the Russian version UDK of the Universal Decimal Classification (RZMat). MSC is a standalone scheme for mathematics, which goes back to a rudimentary scheme developed by the AMS for distributing offprints, was expanded under the name AMS Subject Classification scheme with the aim to classify the entries in Mathematical Reviews and was developed jointly by Zentralblatt MATH and Mathematical Reviews as an international standard in a series of revisions of the AMS Scheme. Applications of mathematics to other sciences are present as particular subject areas.

11-XX	**NUMBER THEORY**
11-00	General reference works (handbooks, dictionaries, bibliographies, etc.)
11-01	Instructional exposition (textbooks, tutorial papers, etc.)
11-02	Research exposition (monographs, survey articles)
11-03	Historical (must also be assigned at least one classification number from Section 01)
11-04	Explicit machine computation and programs (not the theory of computation or programming)
11-06	Proceedings, conferences, collections, etc.
11Axx	**Elementary number theory** {**For analogues in number fields, see 11R04**}
11A05	Multiplicative structure; Euclidean algorithm; greatest common divisors
11A07	Congruences; primitive roots; residue systems
11A15	Power residues, reciprocity
11A25	Arithmetic functions; related numbers; inversion formulas
11A41	Primes
11E72	Galois cohomology of linear algebraic groups [See also 20G10]
11E76	Forms of degree higher than two
11E81	Algebraic theory of quadratic forms; Witt groups and rings [See also 19G12, 19G24]
11E88	Quadratic spaces; Clifford algebras [See also 15A63, 15A66]
11E95	p-adic theory
11E99	None of the above, but in this section
11Fxx	**Discontinuous groups and automorphic forms [See also 11R39, 11S37, 14Gxx, 14Kxx, 22E50, 22E55, 30F35, 32Nxx]** {**For relations with quadratic forms, see 11E45**}
11F03	Modular and automorphic functions
11F06	Structure of modular groups and generalizations; arithmetic groups [See also 20H05, 20H10, 22E40]
11F11	Holomorphic modular forms of integral weight
11F12	Automorphic forms, one variable
11F20	Dedekind eta function, Dedekind sums
11F22	Relationship to Lie algebras and finite simple groups
11F23	Relations with algebraic geometry and topology

A piece of the MSC2010

The current MSC provides 5,606 five-digit classification codes, the first two digits representing the top level of 63 mathematical subject areas and the first three digits leading to the second level of 528 subareas. There always had been relations between the subject areas, or even clusters of current mathematical research, where the subjects were distributed over several established subject areas. The only reasonable way to solve such problems was (and still is) to add cross-references and to extend the tree on the third level (for more information see [9]). The MSC is a living classification scheme. Revisions were made every ten years with a light revision after five years. This is done in close cooperation between Zentralblatt MATH, Mathematical Reviews and the mathematical community.
As mentioned at the beginning, the MSC codes are too coarse to be able to describe the content of a publication with sufficient precision, even though in many cases several MSC codes are assigned to the same publication, distinguishing primary and secondary relevance. Key phrases generally provide more precise information on the subjects a paper is dealing with. But they are not part of a controlled vocabulary and their assignment depends on the judgement of an editor. By taking their verbal explanations MSC codes may be considered as key phrases themselves. Appropriate linking to different parts of a publication of may help to provide tools for making the search more precise. Whatever this may be, editors will need support by automatic processing of the documents to be able to install such facilities. The following should describe some measures to come nearer to a solution of the problem.
A first essential step for the further development of the MSC and and its improved Web integration is the development of a SKOS version, see [10]. This has several advantages:

- SKOS schemes can be processed automatically. SKOS is an W3C standard providing a model and vocabulary for knowledge management systems like thesauri and classification schemes based on the W3C standards like XML, RDF, and OWL. SKOS covers elements (a principal element in SKOS is the 'concept') and attributes as well as main relations in thesauri and classification schemes. XML, RDF, and OWL provide the syntax and the basic models.
- SKOS schemes can be linked with each other. By various 'match' elements within the SKOS vocabulary, the SKOS scheme can be concatenated with other knowledge management systems like a controlled vocabulary. Of course, the specification of the matching relations like the concordances between classification schemes is outside the scope of SKOS and must

be done separately. The matching of different SKOS schemes is a nice property which also can be used for the automatic indexing. This idea is described later.

- SKOS schemes are flexible and extensible. A SKOS scheme can be adapted to the special needs of a community and extended by additional properties. In the case of the MSC this may be an extension by additional relations for the similarity of the MSC classes.
- SKOS schemes can be used to support retrieval by producing special strategies for retrieval like navigation maps.

4. Key phrase extraction and classification using linguistic methods

In the last decade, various methods for text analysis were developed supporting an automatic extraction of key phrases. They are based on statistical methods combined with a linguistic analysis of the text. Such approaches turned out to be successful for the analysis of simple texts, not involving sophisticated expressions. In mathematics like in other sciences the publications are written in a more technical language leading to a lot of complications: The mathematical terminology is very specialized and rich though frequently using common language for the many notions invented by mathematicians. Even the same word is used in different contexts having a different meaning then. In addition to this we find terms and acronyms only used in mathematics. Sometimes subjects may be identified by standardized predicates. The most difficult part is to extract the information on subjects, which is stored in formulae. As a first step these difficulties had been ignored and a first version of an automatic tool for keyword extraction had been developed, which is based on the general methods already available, taking the abstracts or reviews of mathematical publications as the reference for the analysis. An advanced tool for an automatic text analysis should be able to handle both, text and formulae. The big majority of mathematical publications is written in English at present. Hence we may concentrate our efforts on English texts at the beginning, caring about translations into other languages later. Here we are confronted with the common problem of language processing on the mathematical level: the ambiguity in both directions, i.e. different words may describe the same subject, the same word may be used for different subjects. The same applies to formulae: Different formulae may express exactly the same mathematical fact, the same formula may describe different mathematical facts.

Modern mathematics has developed a lot of formalization for the description of mathematical objects, structures, relations, facts and proofs. Hence formulae are a fundamental and indispensable part of mathematical publication. It is impossible to understand a mathematical publication by just ignoring the formulae, and in almost any case it is unreasonable to write one without formulae. formulae allow to express arbitrary complex and structured content in a rather precise form. But they are only of limited use for the automatic extraction of information on the subjects of a publication. Most formulae appear as an intermediate step of an effort, to prove a result or to simplify an equation, and only the final result may be of interest for deriving an information on the subject of the paper. Another difficulty is given by the arbitrary complexity of formulae and various encodings of semantically identical formulae. On the other side, semantically different facts may be expressed by the same encoding as a formula. Hence text analysis based on formulae always will have limitations for the description of the subject of a paper. To combine it with the analysis of the remaining text will improve the result. The best result will be obtained, when a semantic markup for the formulae has been made by the author or an editor, which is possible in principle by encoding the publication in Content MathML. Unfortunately, at present most authors only provide Presentation MathML, generally generated as a conversion from a TeX-encoding.
Text analysis of the plain text of a publication leads to more promising results. Like other sciences mathematics has developed its own set of terms describing mathematical subjects. Several types of terms and environments of terms can be distinguished.

- The mathematical vocabulary: It uses many words from common or technical language and extends this list by notions, which are specific for mathematics. The words taken from common language are labels of well-defined mathematical concepts, like groups, fields, convergence, measure or continuity. In addition to this there exist of words used in mathematics exclusively like algebra, holomorphic, hyperbolicity or polytope. For higher precision combinations of common words and technical notions are used like semicontinuity, subgroup or matrix algebra.
- Named Mathematical Entities (NMRs): NMRs denote objects which are widely used by the mathematical community (de-facto standards). They play an important role in mathematical publications as names or identifiers for formulae, mathematical facts and mathematical objects. They are frequently used in mathematical texts. Often, combinations with

names of persons, which may not be registered in the standard dictionaries of language processing, like the KnasterKuratowskiMazurkiewicz lemma, Fubini's theorem, being almost the same as the Cavalieri principle, or the Erdös number, are named entities in mathematics. Generally more structured subject information is linked with NMRs like definitions, mathematical objects, methods and relations or a whole theory. There may be symbols identifying a NMR or synonyms, substructures, broader terms, terms, similar terms, etc. All this has to be taken into account.

- Acronyms: Names of mathematical objects are often described by longer text phrases. Typically, authors defines acronyms as abbreviations for frequently used phrases. Often, acronyms will be introduced as abbreviations by extracting some special characters from a phrase. Acronyms are tagged by a special spelling (one or more capital letters, also inside a term). It's is good practice that the acronyms used in a paper are expanded at its beginning or end. Acronyms can be ambiguous. For content analysis, it is important that acronyms together with the corresponding part of the given text phrases are interpreted, when a unique text expansion is not available.
- Combinations of phrases: Though also sentences in mathematical texts have the usual distinction of subject, predicate and object, these entries may be quite complicated. For example, a lot constructs of noun phrases appear, the simplest ones given by combinations like '... of ' or ' by '. They have to be considered as a unit for the text analysis.
- Segmentation: Mathematical publications generally have a well-structured presentation. Structural elements which could be distinguished easily are headed by axiom, definition, theorem, lemma, propositions, corollary, proof, example, remarks etc. Also the abstract of an article and the bibliography are typical segments. The type of segment where a word appears has relevance for the semantic of a publication. The segmentation is relevant for the content analysis of complete publications.

In the following we restrict our considerations to the analysis of reviews and abstracts of mathematical articles. This may be justified by the following arguments: As explained above, reviews and abstracts (should) describe the main content of a publication in a condensed form. Hence, though covering a selection of terms only, abstracts and reviews are a reasonable basis for the development of an automatic tool for the key phrase extraction. Problems caused by the big amount of terms provided with the full text of a publication may be handled by a later

release of such a tool. But, one main reservation should be kept in mind. The selection given by an abstract or review can be considered as an initial intellectual phrase extraction and may turn out as insufficient if one looks at the complete text of a paper. In the case of reviews this is not an exception only. Therefore the subject editors at Zentralblatt MATH have to spend additional work to provide a more complete selection of key phrases. Reviews or abstracts contain formulae only in rare cases and in such cases they are restricted to the main mathematical objects the publication is dealing with. This leads to two contrary cases. The formula may just be a supplement to the subject information available in the abstract or review, which is the lucky case for automatic phrase extraction, or the reader should see from the formula what the paper is about, which is the bad case. The full text of the majority of mathematical publications is not available for zbMATH (and any other external party trying to establish an integrated semantic search) for text analysis, whatever the method may be. There is the commercial barrier at the big publishers, who do not open the access to the full text for such purposes and are content with their internal search engines, though the quality of these engines is far from satisfying the requirements formulated in this paper. And there is the other barrier, that a big portion of mathematical papers is not available in fully searchable digital form up to now. These are PDFs without providing access to the source file encoding the paper or scanned images, where a text OCR has to be applied to make the text without formulae searchable. This is not a reasonable basis for developing a tool for a comprehensive text analysis. For us it was natural (and also successful) to start the development of automatic tools for the key phrase extraction with support from computational linguistics. Computational linguistics has the general aim to make natural language texts machine-processable. For this purpose the structure and the rules of the common version of a language like English (syntax, grammars, different types of sentences) were investigated. Dictionaries were developed capturing as many words of (the common version of) a language as reasonable. Of course, there are a lot of structural differences between different languages. Chinese and other Asian languages like Thai, Vietnamese or Khmer language have other characteristics than the European language families like English. Hence, a linguistic analysis requires an adaption of the methods to the specific structure of a language, which is not a big problem in the case of English. Further specification has to take the mathematical ontology into account.

The approach presented in this paper starts with an analysis of reviews or abstracts, which generally consist of a sequence of sentences. A segmentation exhibiting theorems in an explicit form is rare and does not contribute to the semantics of the abstract or review. Hence changing to linguistic terminology, our first step is the tokenization of the abstracts or reviews.
Tokenization: Tokens in English texts are separated by blanks. Blanks are also the decisive criterion for the tokenization of reviews and abstracts. formulae in mathematical texts will be considered as composite tokens, ignoring the blanks in a formula. A further analysis of the formulae will be delegated to later parts of this paper. Further composite tokens in mathematics are combinations of names, common English words, etc. as explained above. Different spellings are used by the community like 'fixed-point' versus 'fixed point', where one token has to be identified with a composition of two tokens, or 'quasiconvexity' versus 'quasi-convexity', where two different tokens have to be identified. A morphological analysis and development of special dictionaries of morphological forms can be used for unification or normalization of the different spellings of the same token.
One of the basic methods in computational linguistics is the type classification of the tokens of a language, known as Part-Of-Speach (POS) tagging: There exist different tag schemes for the English language. Some of the most commonly used schemes are the Penn Treebank scheme with 45 classes the C5 tagset with 61 tags, and the Brown tagset with 87 elements. Here we use the Penn Treebank scheme [11] which has, e.g. , the folllowing tags for nouns: NN (noun), NNS (noun, singular or mass), NNP (noun, plural), NNPOS (proper noun, singular), NNPOP (proper noun, plural). To determine the tags, a lookup in dictionaries is started. The underlying dictionary for our approach is the Brown corpus, a list of more than 1,000,000 English words from 5,000 written texts from different genres. The tokens in the dictionary also include the tags about their type.
Here are some possible difficulties:

- The tagging generally is not unique, a token can be an element of different classes.
- No dictionary is complete. That means that not all potential tokens are in the dictionary.

In these cases, the adequate tag of a token in a sentence must be deduced from the context. This can be done on the basis of rules and/or by the application of stochastic methods, where the 'and' stands for modern methods. The most popular method on the stochastic part is given by the Hidden Markov Models (HMM) to detect the (hidden, because not observable) POS tags. The Dynamic

Programming approach, especially the Viterbi algorithm, is very popular to calculate the adequate tag of a token in a phrase. For the POS-tagging Open-Source software, see [12], is available. The existing Open Source software is used as a starting point for the development of advanced software for the content analysis.
Mathematical formulae are outside the scope of computational linguistics for English as well as for other languages. Hence, formulae are handled by a special approach. By deleting control characters and numbers, an artificial hash for each formula is created. The original formulae are stored separately which allows a special analysis of the mathematical formulae.
The POS tags are an important resource for further linguistic investigations. Grammatical relations and context-free grammars are concepts to formalize relations. For example, the subject-predicate-object rule for English can be used to identify word groups or phrases. As the name 'noun phrase' indicates, it requires an a-priori identification of the nouns and the classes of the tokens in the environment.
Context free grammars (CFGs), also called phrase-structure grammars, are schemes and models for structuring natural languages. CFGs cover a set of rules how the elements of a language can be grouped, e.g., by defining characteristic patterns of noun phrases consisting of elements from the Penn Treebank Tag-set. Again, a dictionary covering tokens and additional grammatical attributes is needed to complete a CFG.
Noun phrases: Noun phrases are the most relevant phrases in mathematical publications for executing a content analysis. Noun phrases can have different roles in a sentence, e.g., as the subject or as the object of a sentence. Noun phrases can have different structures, like a sequence of some nouns. Noun phrases can be arbitrarily long. We have defined a set of characteristic types of noun phrases in mathematical texts: They are defined by sequences of POS tags. Also mathematical formulae inside a noun phrase have to be considered. An important secondary problem is to identify the leading noun of a phrase (the central noun) and to identify the complete relevant phrase, leaving out irrelevant parts. Up to now, the length of key phrases has seven tokens as upper bound.
The extracted candidates for key phrases are the preliminary material for establishing a controlled vocabulary for mathematics. Such a vocabulary consists of the named mathematics entities enriched by the important mathematical key phrases of publications, which up to now are not part of the nomenclature. Controlled vocabularies may be used as a base for the development of thesauri and ontologies, which can be created by adding relations between the elements of a

controlled vocabulary. Languages are subject to changes. Hence controlled vocabularies have to be kept as dynamic objects. As already mentioned, at present there is no controlled vocabulary for mathematics. One reason for this is that the number of candidates is very huge. M. Hazewinkel, see [13], has proposed 120,000 items as an estimate for the size of such a vocabulary. Our experience is that this estimate is out of date already. The currently available automatic extraction tool has led to some millions of candidates for a controlled vocabulary. The size of the vocabulary is one of the barriers to create and maintain a controlled vocabulary without machine support.

Here we try to realize a semi-automatic approach. The extracted key phrases from the reviews and abstracts are used as candidates for the controlled vocabulary. They may be related with other resources, like the MSC codes or entries in the Encyclopedia of Mathematics or the mathematical part of the Wikipedia. This is done for each MSC class enabling a later application to automatic indexing. The frequency of the occurrence of a key phrase is taken as one measure for its relevance. The resulting relations will be checked intellectually by experts. A periodic iteration of this approach guarantees the actuality of the controlled vocabulary. Related key phrases provide an important enhancement of the MSC codes, because they represent a comprehensive explanation of what the code is about.

The implementation of an automatic classification on the basis of a controlled vocabulary will work as follows: To each MSC code its part of the controlled vocabulary is assigned. The controlled vocabulary for the MSC code is transformed into a SKOS scheme. Then, the SKOS scheme of the MSC is matched with the SKOS scheme of the vocabulary. Methods for automatic text classification base on mathematical models and methods. Usually, text classification starts with the vector space model which provides a representation of a document by a vector. More in detail, the document will be split in n different tokens. The n-th component of the vector is the frequency of the n-th token. Instead of single tokens also phrases of tokens may be used. The extracted text phrases of a document and the phrases of the controlled vocabulary of each MSC class can be directly used to compose vectors, which are used then to determine the degree of similarity between the vectors for the key phrases and the controlled vocabulary. Different metrics and approaches are available for this. Most popular approaches are the k-nearest neighbours method (based on the Euclidean metric), the naive Bayes method (a stochastic approach), and the Support-Vector-Machines (SVM) approach basing on a geometric concept called the separation by hyperplanes.

5. Formula analysis

There are a lot of remaining problems like the normalization and standardization of key phrases, which are given in the document in different forms. The last one we will address in this paper is the development of specific methods for the analysis of mathematical formulae. As mentioned above the analysis of mathematical formulae is more complex than that of texts. But, also the language of mathematical formulae has a lot of conventions and standards for basic symbols. At first some remarks about formulae in zbMATH should be made:

- It is easy to detect the formulae in TeX-encoded texts, because they start and end with $-signs. They will be inside the corresponding MathML-namespace elements in MathML encoded text. If we take this as characteristic of mathematical formulae in the database entries of zbMATH, then zbMATH contains with almost 10,000,000 formulae.
- If we use the size of the tokens as a rough measure of the complexity of a formula, most formulae in zbMATH are simple symbols, i.e., single characters, which denote a mathematical object. The frequency of formulae in reviews or abstracts is different for different mathematical subject areas. For example, the frequency of formulae in reviews or abstracts on mathematical applications is lower than that for pure mathematics.
- The same of formula may have different encodings.

In principle, the development of formula analysis can be done in the same way as for text analysis. The segmentation of formulae is a first important step. This is more difficult than that for English texts, because we have no standard separator to split complex formulae. Moreover, single mathematical formula entities can be rather complex, like definite and indefinite integrals. The tokens of a complex formula may be combined in different ways. XML provides a well-defined structure of documents. TeX encoded formulae can be converted automatically into Presentation MathML and also into Content MathML as far as the original encoding provides a unique semantic interpretation. Hence, XML allows to identify the mathematical tokens of a formula. But, the encoding in Presentation MathML provides only a very rough semantic classification of the elements as mathematical identifiers for constants, variables, or operators.

A finer analysis of mathematical formulae is more complicated than of texts. The formula may include free variables. For example, single character formulae like A or n are used as notations for a type of mathematical objects. Other single character formulae like P allow an more specific interpretation as subject when talking about in probability. The use of fonts (bold, cursive, Gothic, etc.) is

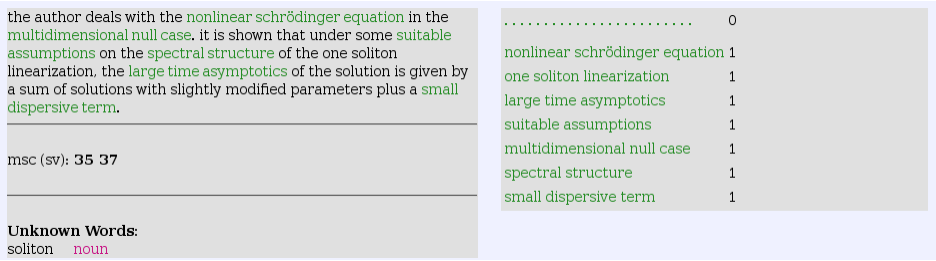
not standardized. Nevertheless, for a lot of named mathematical formula entities standard notations are common practice. Several named mathematical formula entities are characteristic for special subject areas. This may be given by an MSC code. Hence, the semantic background of a formula easily can be identified in this context.

- The development and maintenance of comprehensive catalogue of mathematical formulae is very expensive. A first approach may be given by the concept of the content dictionaries of the OpenMATH enabling the processing of formulae in arbitrary systems. Here we pursue another goal, namely to develop a concept which helps us to identify the tokens of a mathematical formula and make them and the complete formula searchable.
- For this purpose we need a dictionary of mathematical tokens. From segmentation, we get a first list of possible tokens in formulae. But they are of different value for the content analysis. For example, a sentence like 'Let A, B be matrices ...' is only of interest for a combined analysis of formulae and text.
- We also need an analogue for a standardized classification of mathematical formulae like the POS tagset for English language tokens. We need a dictionary listing the possible meanings (as a measure for the ambiguity of a token), matching tables for synonyms, and context information. The context information may be quite general like an MSC code of the publication or a reference. It may be very specific like the left-hand or right-hand side of a formula.
- A grammar for mathematical formulae would be useful. Here a grammar is understood as a set of general rules for the interpretation of the structure of mathematical formulae. This could be used for a deeper analysis of MathML formulae going beyond the path analysis. There are some activities to develop such a grammar, see [14]. Such an activity should be coordinated with existing activities in for mathematical knowledge management like Planet Math, OpenMath etc.

6. The prototype

A first prototype for the text analysis has been developed. The result offers the following features: an extraction of candidates for key phrases, a list of unknown words including the proposed POS tags, a proposal for a MSC classification. At present, the classification is restricted to the subject area level of the MSC).

The following snapshot should give a more detailed impression of the prototype.



The user interface of the prototype for key phrase extraction and classification

On the left-hand side, the input box, which contains the original text (review or abstract), is located. On the right-hand side, the extracted phrases are listed. These phrases are also highlighted in the original text to enable intellectual control by human beings. The left-hand side is completed by the proposal of the MSC subject area and a list of 'unknown' tokens, which are not in the Brown corpus, together with a proposed POS tag.
Two classifiers, the naive Bayes classifier (nv) and the Support Vector Machine classifier (sv), were used to calculate the classification. The selection of key phrases and the correctness of POS tags for the unknown tokens can be steered by check boxes. Unknown words with correct POS tags will be added to the dictionaries. The tokens are weighted by their frequencies.

7. Conclusions and Outlook

Because only the first two digits of an MSC code are provided by the prototype introduced above, this may be considered as a weak solution of the general problem, to establish an advanced semantic search. But this already leads to a first support of the editorial procedures for zbMATH, and it also may be used by librarians, who are not interested in refined classifications of their holdings. The prototype also is a good motivation to spend more efforts to obtain an advanced tool for automatic indexing applying the methods described in this article. The basis for the application of an advanced tool should be the full texts of the publications though the abstracts and reviews may be sufficient for the development of a first controlled vocabulary. As said at the beginning, mathematical literature is available today in digital form, and the best solution would be to have searchable versions for all publication and free access to their presentations for everybody.

This is the dream of the WDML (World Digital Mathematics Library). A comprehensive digital library for the mathematical literature should be distributed, scalable and flexible so that all providers of mathematical literature may take part in the enterprise and all people interested in mathematics can access their publication of interest.

Common standards and efficient methods for content analysis are essential for the quality of such a digital mathematics library. At present we have to be content with partial offers like the ElibM [15] (Electronic Library in Mathematics) in EMIS, hosted at Zentralblatt MATH and representing the largest repository of open access journals in mathematics, and the various national activities offering repositories of digitally born or retro-digitized articles like NUMDAM [16], DML-CZ, ERAM, RusDML, MathNetRu and others. EuDML [17] was a project build up a distributed digital library integrating the mayor European open access providers. Methods for an efficient machine-based content analysis where a topic of high priority for all these projects.

References

[1] Johann Ephraim Scheibel, Einleitung zur mathematischen Bücherkenntnis. (Band 1) Breslau: Meyer (1769, 1775)

[2] Friedrich Wilhelm Murhard, Litteratur der mathematischen Wissenschaften. Leipzig: Breitkopf und Härtel (1797)

[3] Journal für die reine und angewandte Mathematik (Crelle Journal) http://www.degruyter.com/view/j/crll

[4] Felix Müller, Carl Ohrtmann, Jahrbuch über die Fortschritte der Mathematik. Band 1. Berlin: G. Reimer. S. I-IV (1871)

[5] Georg Hermann Valentin, Die Vorarbeiten für die allgemeine mathematische Bibliographie. Bibliotheca Mathematica 1 (3), S. 237-245 (1900)

[6] Felix Müller, Zur Frage der Begründung einer mathematischen Zentralbibliothek. Bibliotheca Mathematica 4 (3), S. 389-391 (1903)

[7] zbMATH (Zentralblatt für Mathematik) http://www.zentralblatt-math.org/

[8] Mathematical Subject Classification http://msc2010.org

[9] Patrick Ion, Wolfram Sperber, MSC2010 in SKOS, EMS Newsletter No. 84, 2012, 55-57

[10] MSC/SKOS http://msc2010.org/resources/MSC/2010/MSC2010

[11] The University of Pennsylvania (Penn) Treebank Tag-set, http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

[12] http://nlp.stanford.edu/software/tagger.shtml#Download

[13] Michiel Hazewinkel, Enriched thesauri and their uses in information retrieval and storage, 1996, E.R.C.I.M.

[14] Deyan Ginev, The structure of mathematical expressions, Master Thesis, 2011, Jacobs University Bremen, http://kwarc.info/people/dginev/publications/DeyanGinev_MScThesis.pdf

[15] ElibM, http://www.emis.de/elibm/index.html
[16] NUMDAM, http://www.numdam.org
[17] EuDML, https://eudml.org/

Received June 1, 2013
Revised version received June 25, 2013

FIZ Karlsruhe, Zentralblatt Math, Fanklinstr. 11, D-10587 Berlin
E-mail address: wolfram@zentralblatt-math.org

Department of Mathematics, TU Berlin, Str. des 17. Juni 136, 10623 Berlin
E-mail address: zblwegner@googlemail.com