New Developments in Electronic Publishing
AMS/SMM Special Session, Houston, May 2004
ECM4 Satellite Conference, Stockholm, June 2004
p. 9-18.

Searching Mathematics on the Web: State of the Art and Future Developments

Andrea Asperti **Stefano Zacchiroli**

Department of Computer Science, University of Bologna
Mura Anteo Zamboni, 7 – 40127 Bologna, Italy
e-mail: asperti@cs.unibo.it zacchiro@cs.unibo.it

Abstract. A huge amount of mathematical knowledge is nowadays available on the World Wide Web. Many different solutions and technologies for searching that knowledge have been developed as well. We present the state of the art of searching mathematics on the web, giving some insight on future developments in this area.

1. Introduction

The World Wide Web has become one of the main resources used by mathematicians in every day work. Its usefulness is not limited to browsing fellow researchers, university, or research projects web pages. A full range of mathematical services are available as well on the web, ranging from electronic libraries of mathematics to communities of distributed agents, implemented as web services, able to cooperate in order to solve a given mathematical problem.

Searching such a huge amount of mathematical sources is a particularly complex problem, given the variety of different users with completely different needs, and the heterogeneity of the mathematical information and its possible encodings.

The different kind of queries may be roughly categorized in three main groups:

Bibliographic searches: this is the most traditional kind of query, aimed at retrieving a document given its author, title, date of publication, a list of keywords or similar information. A typical query could be e.g. *give me a listing of all articles written by Karl Weierstrass on the subject of analytic functions.*

Mathematical services: in this case the user in typically interested to *solve* a problem with the help of some mathematical tool, or a combination of them. A typical query in this context could be the request for a web service able to establish the primality of a number given as input by the user.

Content based searches: the third and probably most ambitious category of queries are those based on the mathematical content of the information (opposed to its textual representation). These queries are aimed at a very fine-grained analysis of the repository, looking e.g. for all documents stating something about the expression $cos(z) + i\ sin(z)$, where of course z has to be understood as an universally quantified variable whose actual name is thus irrelevant (try with Google!).

In this paper we discuss in more detail the previous categories of queries, presenting the state of the art and the main research directions. A particular attention will be devoted to *mathematical services* and, especially, *content based searches* being the most innovative and peculiar kind of queries for mathematical repositories.

2. Bibliographic searches

Most part of printed human knowledge available all over the world is stocked inside libraries. The problem of how to search those libraries in an efficient manner has been traditionally solved by the combined use of metadata (data about the documents themselves) and indexes. In this case, mathematical documents do not substantially differ from other kinds of documents and standard knowledge management and indexing techniques can be profitably applied. From each document of the library a set of metadata is extracted and several orthogonal indexes are created on top of them referencing the physical locations of the actual documents. Searches performed using that indexes are called *bibliographic searches*.

Metadata could include information about document title, author, editor, classification and so on. The classification field is of particular interest for searches since it defines a taxonomy over human knowledge: related documents are likely to share a common classification.

Since the beginning, several classifications have been developed by librarians, the most widespread being the Dewey Decimal Classification [1] in which "Natural sciences & mathematics" have been assigned decimal number 500. Since Dewey's is too lax for properly classifying mathematical documents, other classifications have been developed by mathematicians, the most widespread being MSC 2000 [2] – Mathematical Subject Classification Scheme – maintained by the American Mathematical Society. In this classification for example 35E15 is a subtopic of partial differential equation (35xxx) with constant coefficients (35Exx), documents with that classification will be about the initial value problem (35E15).

The arrival of the web era has increased accessibility of mathematical libraries and improved the expressivity of queries, but has not (yet?) radically changed the way in which bibliographic searches are performed.

Zentralblatt MATH [3], coordinated by FIZ Karlsruhe, is the most prominent project in this area being the longest running indexing and abstracting service in pure and applied mathematics available on the web. It classifies more than 2,000,000 entries accordingly to MSC 2000. Searches are possible via various fields, like author, title, document type, MSC classification, year of publication and so on. Boolean combinations of the various fields for fine grained searches are possible as well.

For each results title, abstract and all metadata information associated to the document are shown and other interesting actions are possible like browsing related documents (same

MSC 2000 classification) and on-line ordering of the printed document.

While Zentralblatt is the most relevant indexing and abstracting services, other on-line databases of mathematical documents are available offering, on a smaller scale, similar classification services. Just to name a few of them: MathSciNet [4] by the American Mathematical Society and the Electronic Research Archive for Mathematics (ERAM) [5].

2.1. The European project Euler

A big improvement in the accessibility of all these electronic libraries have been induced by the European based Euler project [6]. On the behalf of this project a web based gateway, with searching capabilities really similar to those of Zentralblatt MATH, has been developed. Using Euler the user has access to the catalogues and repositories of mathematical documents of participating institutions, while the latter keep control over creation and maintenance of their data.

Euler is the state of the art of mathematical bibliographic searches on the web: a portal offering unified access to documents from Zentralblatt, the CWI database of the Dutch national research center of mathematics [7], ERAM and many more institutions. Once an entry is found, access to the electronic or printed version of the document is mediated via the web site of the document owner.

Euler is currently supported by the European Community as a take-up Project (n.IST-2000-29445), based on the achievements of the successfully completed EULER project (FP4 "Telematics for Libraries" project LB-5609).

2.2. Key phrases and information clouds

To manage huge scientific archive metadata are essential, in particular key phrases which should come from a large standardized (controlled) and updateable list. A major problem here is that a perfectly good key phrase for a given chunk of text may very well simply not occur there (or be so linguistically disguised that it cannot actually be recognized). Scientists get around this by looking at the surrounding text.

This is the idea of an *identification cloud* which in its simplest form is just a list of words (possibly with weights) that is attached to a standard key phrase and that are likely to occur in texts dealing with the concepts embodied by that key phrase. The concept was introduced in the ongoing project Trial Solution (IST-1999-11397) [8] with promising results.

3. Mathematical services

In the last years several research efforts have been made for the standardization and the deployment of web services [9]. Fitting properly in the semantic web [10] framework, web services are software systems designed to support machine to machine interaction over a network (usually the Internet), typically exposing a programming interface based on exchange of XML documents [11].

3.1. The W3C standardization effort

The standardization activity of the World Wide Web Consortium in the area of Web Services is articulated in 5 working groups: Web Services Architecture Working Group, XML Protocol Working Group, Web Services Description Working Group, Web Services Choreography Working Group and an auxiliary Coordination Group. The most relevant ones are:

Web Services Description Working Group: This group[1] is chartered to design an XML based language that should be able to describe a web service *interface*. This task includes also the design of web service *messages*, *message exchange patterns* and *protocol bindings*.

The group has already released, among other documents, a working draft of WSDL (Web Services Description Language) 2.0 [12] and an additional document which describes bindings of this language to other existing technologies like SOAP, HTTP GET and POST methods, MIME [13].

Web Services Choreography Working: This "young" group[2], started in January 2003, is chartered to design a language that is able to describe choreographies of web services. Intend meaning of a *Web Service Choreography* is some kind of interaction between web services.

One possible usage of a choreographies is the creation of complex web services simply composing simple web services as we compose functions in math.

3.2. The Monet project IST-2001-34145

Web service technologies could be really effective in solving long standing problems of inter-application communication. Still, W3C standards provide only a framework in which this problem could be solved and do not instantiate the technologies to specific fields of interest. The road of standardizing and deploying web service technologies for the special needs of mathematicians has been took by the European Community funded Monet project [14, 15]. Aim of the project, recently completed, was the development of a framework in which mathematical web services can describe their capabilities in as much detail as is necessary to allow a sophisticated software agent to select a suitable service based on an analysis of the characteristics of a user's problem.

Using standards and technologies developed by the Monet project it is possible to implement what we call a mathematical *functional search*, that is finding on the network a web service able to resolve a given mathematical *problem*[3].

Characteristics of mathematical web services in Monet are described in the XML based language MSDL [16] (Mathematical Service Description Language). An MSDL document is composed of several parts: classification, implementation details, service interface and binding descriptions, broker interface and service metadata. All these data could be used for querying an *Instance Store* (IS) for available services.

[1] `http://www.w3.org/2002/ws/desc/`

[2] `http://www.w3.org/2002/ws/chor/`

[3] Monet takes care of several other aspects of mathematical web services like client-broker architecture, publishing and discovery of services, planning, orchestration and so on. We will focus our discussion on the discovery part of Monet

From the point of view of the mathematician, the most interesting part is the classification, a specification of *what* the service does. Classification is done on several axis: at each service several classifications could be applied and each of them could be used in user queries.

A first classification is done giving a description of the *mathematical problems* the service is able to solve. Descriptions include problem inputs, output and pre/post conditions. For example, minimization of a multivariate function over the real numbers could be described as follows:

Input: $F : \mathcal{R}^n \to \mathcal{R}$

Output: 1. $x \in \mathcal{R}^n$; 2. $m \in \mathcal{R}$

Post-conditions: 1. $F(x) = m$; 2. $\nexists y \in \mathcal{R} \mid F(y) < m$

MSDL mandates the XML format of such problem descriptions, but not the format used to describe mathematical formulae like the above one. They are usually encoded in OpenMath [17] fragments, but they can be MathML content [18] fragments as well. Additionally, each problem belongs to a specific class of problems, for example `definite_integration` for problems related to definite integration.

Other interesting ways of classifying mathematical web services in Monet are by the means of *standardized taxonomies* like GAMS (Guide of Available Mathematical Software) [19] from NIST, and accordingly *algorithms* implemented by the service.

Once a web service has been described in a MSDL document, it can be registered to one or more ISs and found by users via queries.

The simplest form of queries is just a MSDL document used as a template: user fill some fields and leave others empty. For example if we are looking for any service able to compute the eigenvalues of a matrix, we have simply to create a fake MSDL description of a service in the GAMS class D4a ("Ordinary eigenvalue problems"). Unfortunately, Monet has not yet developed any user friendly interface for writing MSDL documents, and for the moment a user has to manually write the XML document by hand.

One of the main contribution of Monet is a set of *ontologies* which could be used by the IS. One ontology is for example defined over the GAMS taxonomy so that the query engine is able, using an external reasoner over ontologies, to return not only services in GAMS class D4a for the above example query, but also services in all subclasses, for example D4a5 ("Ordinary eigenvalue problems on tridiagonal matrixes").

However, the most useful way of querying an Instance Store is not the basic template query, but rather the queries *via user input*. An IS could for example queried with the following input (properly encoded in OpenMath):

$$\int_{0.0}^{1.0} \sin(x)\,\mathrm{d}x.$$

We may now see an interesting use of one of the Monet ontologies: the problem ontology. In this ontology each class of mathematical problems is associated to an `openmath_head` property which reference an OpenMath symbol likely to be found in problem instances. For `definite_integration` this symbol is `defint`, the definite integral symbol $\int_y^x$. Due to the tree structure of OpenMath formulae the IS could easily match the above user input with

problems of the `definite_integration` class and retrieve all the services computing definite integral.

On user request, the IS could do even more instantiating service inputs from user input and return to the user directly the result of the integral computation.

Monet standards are really promising, but in order for them to be useful we still have to wait for deployment and actual use of related technologies by widespread mathematical software packages.

4. Content based searches

The World Wide Web is already the largest resource of mathematical knowledge, however almost all mathematical documents available on the web are marked up only for presentation, preventing any attempt of automation, interoperability, transformation or processing. The long-term goal, in this area, is to overcome these limitations passing from a machine-readable to a machine-understandable representation of the information. In its deepest semantical sense, this means moving towards a real formalizations of the mathematical content, that is an essential prerequisite for most kind of automatic processing.

The interest around content (or semantic) based functionalities has been growing in recent years (see e.g. [20, 21, 22, 23, 24]). For instance, [23] provides a quite paradigmatic case study. The problem is to achieve an automatic classification of the differential equations appearing in some mathematical document according to a given set of criteria (comprising e.g. the order, whether is ordinary or partial, whether it is linear, homogeneous and so on). This implies a reconstruction of the semantics of the formula, possibly its transformation (e.g. its normalization to some kind of normal form), and finally its analysis.

The general problems are essentially of three different kinds:

1. provide tools which help to automatically reconstruct the mathematical content of mathematical expressions (especially meant for legacy documents), and/or help to assist a direct content-based authoring of new documents,
2. define and maintain well documented *content dictionaries*[4], fixing the ontology of relevant fragments of the current mathematical notation,
3. developing techniques, languages and tools supporting and exploiting innovative content based functionalities.

4.1. The MoWGLI project IST-2001-33562

The goal of exploiting content based techniques for mathematical knowledge management has been explicitly addressed by the European Project IST-2001-33562 MoWGLI[5] (ending in March 2005). MoWGLI was also intentionally conceived as a major test-bench for XML-technology (DOM, Stylesheets, MathML, SVG, RDF, XMLQuery, etc.) whose extensive use and testing has been a major leitmotif of the project.

The main technological issues dealt within MoWGLI have been *rendering* (in particular, rendering engines for MathML), *web publishing* (mostly done on the fly from XML sources

[4] We use here the terminology of the OpenMath Consortium [17]

[5] `http://mowgli.cs.unibo.it`

via XSLT transformations), and *searching* (via a rich set of metadata automatically extracted from the content description of the information).

Developing and testing these tools required the possibility of having at our disposal, and as soon as possible, large collections of documents encoded with semantic markup. One strategy which has been pursued has been to develop a tool (called *Hermes*) supporting a LaTeX-based authoring mode for mathematical articles. A more rapid way to get meaningful repositories of fully structured mathematical knowledge was by exporting them from the available libraries of Logical Frameworks and Proof Assistants. In particular the library of the Coq [25] proof assistant developed at INRIA has been successfully exported into XML and is currently searchable and browsable by means of MoWGLI's technology at `http://helm.cs.unibo.it/`.

4.1.1. Searching via metadata

One of the main achievement of MoWGLI is the feasibility of performing sophisticated queries via a restricted set of metadata automatically extracted from the structured representation of the information. The structural content is essential since it allows to consider the *position* of items in the document as a viable metadata. Although canonical indexing techniques like tree automata, discrimination trees [26], substitution trees [27] or context trees [28] can be profitably used in order to solve specific matching problems, the metadata approach provides a degree of flexibility and modularity that goes beyond (and is orthogonal to) all previous approaches.

Even in the case of iterated searching finalized to automatic proving (that is one of the most widely investigated subject in the literature) the cost of retrieving and applying all matching statements is in fact neglectable with respect to the exponential growth of the search space along different branches. Finding good heuristics to control the branching factor and having a simple mechanism to filter out unwanted solutions is thus more relevant than implementing an efficient matching algorithm. Metadata provides such a simple, effective and absolutely flexible mechanism.

In particular, MoWGLI's metadata for mathematical expressions have the general shape

$$Ref(source, target, position)$$

stating that *source* contains a reference to *target* at the specified *position*. By suitably describing such a position we may reach the same degree of granularity of a substitution tree (where each variable marks indeed a different position), thus providing a complete description of the term. However, a very minimal set of "interesting" positions, that, essentially just discriminates among hypothesis and conclusions, and root operators from inner ones, turns out to be already extremely effective.

A prototype implementation of a content based search engine called Moogle has been developed at the University of Bologna and is available on the web at `http://helm.cs.unibo.it`. The engine is used to search the library of the Coq Proof Assistant [25], composed of about 40000 theorems in various fields of mathematics. Moogle supports 4 kinds of queries (see also [29, 30]):

locate: this query is used to retrieve a mathematical item (theorem, definition, lemma, ...) from its short name.

match: match allows to retrieve a mathematical item from a pattern provided by the user (possibly containing wild cards). For instance, we may retrieve a proof of a statement contained in the repository from the statement itself.

elim: Coq is a logical framework based on the Calculus of Inductive Constructions, and induction is the main logical tool. Several induction principles (traditionally called elimination principles in this context) could be associated to a given datatype. The elim query takes in input a datatype and gives back the list of its elimination principles.

hint: hint is probably the most sophisticated query. It takes in input a (closed) statement and gives back a list of theorems of the repository which can be applied to the conclusion of the statement in the attempt of proving it in a backward fashion.

4.1.2. Hermes

The main problem of semantic-based techniques is still the actual authoring cost of semantically enriched documents.

A second major technical achievement of MoWGLI is the development of an authoring tool for scientific documents (Hermes[6]) supporting manual and automatic generation of MathML content, and automatic generation of MathML presentation, enabling an author to add and recover semantic depth and clarity to LaTeX written documents.

The prototype implementation of Hermes has the following components:

- A set of helper LaTeX macros, which allows the author to disambiguate the meaning of the mathematical expressions he writes, while allowing some choices for the presentation; this set is included by the author in the originally written LaTeX document. A LaTeX run on the macro-enriched document will output a "semantic DVI" file (a DVI file containing "special" annotations of various combinations of graphical and non-graphical symbols in the source).
- A scanner which extracts from the resulting DVI file the semantic tokens seeded by the macro collection above and sends them to the parser below.
- A parser creating the final content-oriented XML representation.
- An XSLT stylesheet, transforming the content-oriented XML document into a renderable, cross-referenced, document.

[6] `http://psyx.org/romeo/hermes/`

References

[1] *Dewey Decimal Classification*. http://www.oclc.org/dewey/

[2] *Mathematical Subject Classification*. American Mathematical Society. http://www.ams.org/msc

[3] *Zentralblatt MATH*. http://www.emis.de/ZMATH/

[4] *MathSciNet, Mathematical Reviews on the Web*. http://www.ams.org/mathscinet

[5] *The Jahrbuch Project Electronic Research Archive for Mathematics* (ERAM). http://www.emis.de/projects/JFM/

[6] EULER – *Your Portal to Mathematics Publications*. http://www.emis.de/projects/EULER/About.html

[7] *Centrum voor Wiskunde en Informatica Library*. http://www.cwi.nl/library/

[8] Hazewinkel, M.: *Statistics of information clouds*. Discussion paper in the framework of the TRIAL SOLUTION project, CWI Amsterdam, 14 Sept. 2001.

[9] *Web Services Activity*. World Wide Web Consortium. http://www.w3.org/2002/ws/

[10] *Semantic Web*. World Wide Web Consortium. http://www.w3.org/2001/sw/

[11] Bray, Tim; Paoli, Jean; Sperberg-McQueen, C. M.; Maler, Eve (eds.): *Extensible Markup Language* (XML) 1.0 (2nd Edition), W3C. Recommendation 2000. http://www.w3.org/TR/2000/REC-xml-20001006

[12] Chinnici, Roberto et al. (eds.): *Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language*. http://www.w3.org/TR/wsdl20/

[13] Haas, Hugo et al. (eds.): *Web Services Description Language (WSDL) Version 2.0 Part 3: Bindings*. http://www.w3.org/TR/wsdl20-bindings/

[14] Caprotti, O.; Dewar, M.; Turi, D.: *Mathematical Service Matching Using Description Logic and OWL*. In: Proceeding of the Third International Conference on Mathematical Knowledge Management, MKM 2004. Bialowieza, Poland. LNCS **3119**.

[15] Caprotti, O.; Davenport, J. H.; Dewar, M.; Padget, J.: *Mathematics on the (Semantic) NET*. In: Proceedings of ESWS 2004, First European Semantic Web Symposium. LNCS **3053**, 213–224.

[16] *Mathematical Service Description Language: Final Version*. The MONET Consortium (IST-2001-34145). Deliverable D14. http://monet.nag.co.uk/cocoon/monet/publicdocs/monet-msdl-final.pdf

[17] *The OpenMath Society, The OpenMath* 2.0 *Standard*. http://www.openmath.org/standard/om20-2004-06-30/omstd20html-0.xml

[18] Carlisle, David et al. (eds.): *Mathematical Markup Language (MathML) Version* 2.0 (Second Edition), W3C, Recommendation 21 October 2003. http://www.w3.org/TR/2003/REC-MathML2-20031021/

[19] *Guide to Available Mathematical Software*. http://gams.nist.gov/

[20] Asperti, A.; Wegner, B.: *An Approach to Machine-Understandable Representation of the Mathematical Information in Digital Documents*. In: Fengshai Bai and Bernd Wegner (eds.): Electronic Information and Communication in Mathematics, LNCS **2730** (2003), 14–23.

[21] Bancerek, G.; Rudnicki, P.: *Information Retrieval in MML.* In: Proceedings of the Second International Conference on Mathematical Knowledge Management, MKM 2003. LNCS **2594**.

[22] Cairns, P.: *Informalising Formal Mathematics: Searching the Mizar Library with Latent Semantics.* In: Proceeding of the Third International Conference on Mathematical Knowledge Management, MKM 2004. Bialowieza, Poland. LNCS **3119**.

[23] Draheim, D.; Neun, W.; Suliman, D.: *Classifying Differential Equations on the Web Statements.* In: Proceeding of the Third International Conference on Mathematical Knowledge Management, MKM 2004. LNCS **3119**.

[24] Bancerek, G.; Urban, J.: *Integrated Semantic Browsing of the Mizar Mathematical Repository.* In: Proceeding of the Third International Conference on Mathematical Knowledge Management, MKM 2004. Bialowieza, Poland. LNCS **3119**.

[25] *The Coq proof-assistant.* `http://coq.inria.fr`

[26] McCune, W.: *Experiments with discrimination tree indexing and path indexing for term retrieval.* Journal of Automated Reasoning **9**(2) (October 1992), 147–167.

[27] Graf, P.: *Substitution Tree Indexing.* Proceedings of the 6th RTA Conference, Springer-Verlag LNCS **914**, 117–131, Kaiserslautern, Germany, April 4–7, 1995.

[28] Ganzinger, H.; Nieuwehuis, R.; Nivela, P.: *Fast Term Indexing with Coded Context Trees.* Journal of Automated Reasoning. To appear.

[29] Guidi, F.; Sacerdoti Coen, C.: *Querying Distributed Digital Libraries of Mathematics.* In: Proceedings of Calculemus 2003, 11th Symposium on the Integration of Symbolic Computation and Mechanized Reasoning. Aracne Editrice.

[30] Asperti, A.; Selmi, M.: *Efficient Retrieval of Mathematical Statements.* In: Proceeding of the Third International Conference on Mathematical Knowledge Management, MKM 2004. Bialowieza, Poland. LNCS **3119**.

Received August 5, 2004