WDS'08 Proceedings of Contributed Papers, Part I, 240–244, 2008. ISBN 978-80-7378-065-4 © MATFYZPRESS

Indexing Mathematical Content Using Full Text Search Engine

J. Mišutka
Charles University, Faculty of Mathematics and Physics, Malostranské nám. 25, 118 00, Praha 1, Czech Republic.

Abstract. Full text search engines are inseparable part of everyday search for information on the WWW. During the last years it became the main resource of mathematical knowledge. Despite the clear importance of a mathematical search engine, this research field had been abandoned until very recently. Although currently available full text search engines can be used on these documents too, they are deficient in almost all cases. Many problems are the result of the mathematical nature. By applying axioms, equal transformations, and by using different notation each formula can be expressed in numerous ways. Ambiguous searches like "sin" or "a" would return documents containing sine function and the English noun sin or documents containing variable a and indefinite article a. Moreover, mathematical operators and special notation cannot be expressed in their query languages. In this paper, we present the current state-of-art of searching in mathematical content. All known mathematical aware full text search engines - MathDex, LeActiveMath, EgoMath - will be described in more detail together with the most applicable non full text approach - MathWebSearch.

Introduction

Most scientific papers are published electronically. Moreover, great efforts have been made to digitise mathematical knowledge in the last years. The number of scientific documents on the WWW is extensive and growing rapidly. Mathematical search engine can be used to find scientific documents containing specific formulae without knowing precise titles or to find scientific documents containing similar formulae.

Although currently available full text search engines can be used on these documents too, they are deficient in many cases. They cannot handle structured mathematical text and mathematical operations. Ambiguous searches like "sin" or "a" would return documents containing sine function and the English noun sin or documents containing variable a and indefinite article a. Mathematical operators and special notation cannot be expressed in their query languages. However, the success of full text search engines has shown that despite missing semantic information satisfactory search results to ambiguous textual queries can be produced. If the search engine can be extended to search for mathematical formulae we can build a fully-fledged mathematical search engine. Uniform style of scientific papers has been exploited by several search engines (Google Scholar[1], CiteSeer[2]) but is applicable only for simple textual searching. Similar techniques used in theorem provers and proof checkers could be applied when focusing on the mathematics. The most important disadvantages of these techniques are the applicability, performance (cannot handle more than few thousands formulae) and zero fault tolerance - similar searching is not possible.

The advantage of using full text search engine to index mathematical content is clear. They can index text documents with little or completely without any document's meta-data and still be very valuable as today's most favourite full text search engines prove. Many mathematical formulae have special titles or hold only in special structures or theories. The possibility of

[1] http://scholar.google.com

[2] http://citeseer.ist.psu.edu

searching in surrounding text can greatly improve the precision of found documents. According to a survey, users of a mathematical search engine find the feasibility of textual search very important [1].

Mathematical Search Engines

There are several possibilities how to categorise search engines which can cope with mathematical content. We will use two main categories: 1) mathematically not aware search engines, 2) mathematically aware search engines. Search engines which can provide better search results to queries about mathematical content than common full text search engines, but on the other hand, cannot handle mathematical operations fall into the first category: 1) Google Scholar, 2) CiteSeer, 3) Whelp [2] (based on search engine for the Helm project) - searching only in provided meta-data. Search engines which can handle mathematical operations or can search for mathematical formulae fall into the second category. We can categorise the mathematically aware search engines according to their approach to: 1) syntactic approach - MathDex, LeActiveMath, and 2) semantic approach - EgoMath, MathWebSearch, MBase [3]. We will not cover MBase because it is not based on a full text search engine and uses pattern matching of the underlying programming language.

MathDex

MathDex[3] [4; 5] is the oldest mathematical aware full text search engine based on Apache Lucene search engine [6]. It went public at the beginning of 2007. The key features are: 1) support for semantically poor documents, 2) accepting different types of mathematical encoding, 3) allows searching on both mathematical notation and text, 4) attempts to match user text search expectations rather than strictly following the query.

All retrieved documents are converted to XHTML+MathML. This search engine can handle semantically poor documents; therefore, it must perform extensive normalisation. Documents are sorted according to their structural and syntactical similarity to the search term. MathDex introduces n-gram matching technique to increase precision. In the index phase not only simple words are indexed but also information about subformulae frequencies is collected. The frequencies of meaningful subformulae can be used to increment the ranking of complex formulae.

Another technique used to increase precision is to separate document into more fields (e.g. title, body) and assign them different weights. Words in the title are ranked higher than words in the body of a document. MathDex stores different parts of an expression as separate fields to allow parallel searching and flexible weighting of matches from different parts of the equation. Each formula is separated into numerator field, superscript field, rows field etc. The input formula is parsed, appropriate fields identified and the query is rewritten to match the subterms in the selected fields. It seems that the probability of retrieving and ranking a document with many formulae is higher than with fewer formulae. MathDex cannot handle mathematical operations nor α-equivalence matching[4].

According to [5], MathDex indexes around 25000 documents from the arXiv, 12000 pages containing mathematics from Wikipedia, approximately 1300 pages from Connexions portal, and around 1000 pages of Wolfram MathWorld.

[3] http://www.mathdex.com:8080/mathdex/search

[4] Alpha-equivalence is a notion of equivalence on terms with binding structure. It captures the notion that the names of bound variables are unimportant; all that matters is the binding structure they induce.

LeActiveMath

LeActiveMath[5] [7] is an intelligent web-based learning mathematical environment. The main goal is to present a learner personalised content based on her previous work, actual knowledge etc. Currently, the system does not provide public content and only subscribers or members can use it. It is based on the Apache Lucene search engine.

It is clear that such a system needs a search engine. The semantic content of mathematical documents is encoded in OMDoc [8; 9]. The indexing phase heavily depends on special OMDoc format which includes semantic information and other meta-data. The applicability on real documents is questionable. However, the use of special OMDoc format can boost the relevance of documents thus making this search engine very helpful in specialised environment. The nature of OMDoc is used to split the content to units called items: theorems, exercises, proofs, definitions, etc. They can be addressed by unique identifiers which help exploiting relations between them. The indexing phase converts the OMDoc formulae into special textual tokens including depth information. The depth information of subformulae is included in the indexed string. The searching phase must convert the input formula to a representation including depth level too. Currently, their solution to this problem is to iterate from 1 to the maximum formula depth in their index database. The problem becomes more significant when searching for complex subformulae. Based on the description of the indexing phase it seems that it has neither support for mathematical operations nor for α-equivalence.

EgoMath

EgoMath [10] is a mathematically aware search engine based on the Egothor v2 full text search engine [11]. The main goal of this search engine is to be applicable for a large collection of real-world documents which do not implicitly contain semantic information. It accepts both Presentation and Content MathML with the focus on Presentation MathML. Documents in the PDF format are converted to Presentation MathML using Infty [12] converter. It supports both the textual search and searching for mathematical formulae. EgoMath represents one formulae not only by one word (complex formulae are represented by an ordered set of words) but by several words (complex formulae are represented by ordered sets of words). Next representation is obtained by applying transformation and generalisation rules to the previous representation. These rules try to reduce the most important disadvantage of the full text index database - static character. Each later representation is generalised, thus allowing more formulae to match the representation. In the searching phase user input is separated into simple textual query and mathematical query. Afterwards, the mathematical query is processed by the same algorithm used in the indexing phase. The algorithm produces N representations which are appended to the simple textual query using the Boolean operator AND. The result are N sequentially executed search queries. Later queries have higher probability becoming hits because each later mathematical representation is more generalised than the previous one. Each mathematical document is separated to mathematical part and textual part. Otherwise ambiguous searches like "sin" or "a" would return documents containing both the sine function and the English noun sin.

The query language is similar to LaTeX. The main disadvantage is that it is still not publicly available because of unfinished UI implementation. Currently, the index database contains 421 documents from the Connexions portal and 1915 documents from arXiv.

[5] http://www.leactivemath.org/

MathWebSearch

MathWebSearch [13][6] is a mathematically aware search engine not based on a full text search engine. The first public version was available at the beginning of 2007. To behave mathematically correct it must heavily depend on the semantic information which can result in significant decrease of precision and recall when operating over documents with little semantic information. As the only search engine it fully supports α-equivalence matching.

MathWebSearch is an example of a non-textual approach, where expressions are parsed into a substitution tree (more commonly used in symbolic math systems such as theorem provers). The result is a tree-like structure with nodes containing substitutions of its parents. The formula is constructed from a root node by applying one or more substitutions. To allow searching for subformulae, it has to add all its subformulae to the index separately. Although MathWebSearch is the most mathematically aware search engine, it must address the same issues as the other mentioned mathematical search engines. On one hand, distinguishing between syntactically same notations but semantically different ($\int f(x)dx$ can mean Riemann but also Lebesgue integral), on the other hand, identifying and grouping syntactically different but semantically equal formulae (${}_nC^k$, C_k^n, $\binom{n}{k}$, $\frac{n!}{k!n!(n-k)!}$). Every search for a formula must traverse a substitution tree which can introduce the problem of performance. The lack of full text search makes the applicability of this search engine very limited.

Currently it can index documents in Content MathML, limited Presentation MathML and OpenMath format. The search engine web front-end is very similar to MathDex but the query language is different. An extended form of MathML is used. According to [13], their index database includes documents from the Connexions portal totalling 3400 articles. The number of terms represented in documents is approximately 53000 (77000 including subterms). They also include an index of the 87000 Content MathML formulae from Wolfram MathWorld.

Conclusion

There are several mathematically aware search engines applicable to larger collection of documents available today. Two of them try to exploit the structure of mathematical formulae from the syntactic point of view, one uses mainly semantic approach and one uses both. One of the challenges of this research field is how to measure the applicability. We think that at this moment, only an exhaustive cross comparison of available search engine can produce useful information. Because this is not available, we make a summary based on the facts described above.

WWW is the biggest repository of scientific documents. The most used document format to publish mathematical papers - PDF - does not directly support description of formula semantics. Therefore, one of the basic requirements for a mathematical search engine to be applicable for WWW is to index semantically poor documents. Another important requirement is to be able to search in the raw text. Currently, it seems that the most applicable search engine for the WWW is EgoMath. There are few portals specialised for mathematical content (e.g. Connections[7], Wolfram MathWorld[8]) which contain description of the formula meaning. When we omit the absence of simple textual queries then MathWebSearch seems the best choice for this type of portals. It seems that neither of these solutions could match a search engine specialised for specific purpose - like LeActiveMath trying to support learning in a specialised environment. Because EgoMath is still not publicly available, the first mathematical search engine - MathDex - with the most heterogeneous index database seems the best choice for current mathematical searching on the WWW.

[6] http://search.mathweb.org/

[7] http://cnx.org/

[8] http://mathworld.wolfram.com

Mathematical searching is currently an active research field with several ongoing projects. One of the most important goals for all mathematical aware search engines is to extend their index databases with portals such as Wikipedia which would make it useful for a larger group of users.

References

Zhao J., Kan M., Theng Y., L.: Math Information Retrieval: User Requirements and Prototype Implementation. To appear in JCDL'08, Pennsylvania (2008)

Asperti A., Guidi F., Sacerdoti Coen C., Tassi E., Zacchiroli S.: A content based mathematical search engine: Whelp. Proceedings of the TYPES 2004, LNCS 3839, Springer Verlag, 17–32 (2004)

Kohlhase M., Franke A.: MBase: Representing knowledge and context for the integration of mathematical software systems. Journal of Symbolic Computation, Special Issue on the Integration of Computer algebra and Deduction Systems, 365–402 (2001)

Miller B., Youssef A.: Technical aspects of the digital library of mathematical functions. Annals of Mathematics and Artificial Intelligence, 121Ũ-136 (2003)

Miner R., Munavalli R.: An approach to mathematical search through query formulation and data normalization. Towards Mechanized Mathematical Assistants, MKM 2007, 342–355 (2007)

Apache Foundation: Lucene Project,
http://lucene.apache.org

Libbrecht P., Melis E.: Methods for access and retrieval of mathematical content in ActiveMath. Proceedings of ICMS 2006, LNAI 4151, Springer Berlin/Heidelberg, 331–342 (2006)

Melis E., Bűdenbender J., Goguadze G., Libbrecht P., Ullrich C.: Knowledge representation and management in Active- Math. Annals of Mathematics and Artificial Intelligence, 47–64 (2003)

Kohlhase M.: OMDoc: Towards an openmath representation of mathematical documents. Seki Report SR-00-02, Fachbereich Informatik, Universität des Saarlandes (2000) http://www.mathweb.org/omdoc

Mišutka, J., Galamboš L.: Mathematical Extension of Full Text Search Engine Indexer. Proceedings of ICTTA'08, IEEE Catalog number CFP08577, Syria, 207–208 (2008)

Egothor v2 search engine,
http://www.egothor.org

Suzuki M., Tamari F., Fukuda R., Uchida S., Kanahori, T.: INFTY - An integrated OCR system for mathematical documents. Proceedings of DocEng, France (2003)

Kohlhase M., Şucan, I. A.: A search engine for mathematical formulae. Proceedings of Artificial Intelligence and Symbolic Computation, AISC'06, LNAI 4120, Springer Verlag, Germany (2006)