Math Information Retrieval: User Requirements and Prototype Implementation

Jin Zhao
Department of Computer Science, School of Computing
National University of Singapore
Singapore, 117590
zhaojin@comp.nus.edu.sg

Min-Yen Kan
Department of Computer Science, School of Computing
National University of Singapore
Singapore, 117590
kanmy@comp.nus.edu.sg

Yin Leng Theng
Division of Information Studies, Wee Kim Wee School of Communication & Information
Nanyang Technological University
Singapore, 637718
tyltheng@ntu.edu.sg

ABSTRACT

We report on the user requirements study and preliminary implementation phases in creating a digital library that indexes and retrieves educational materials on math. We first review the current approaches and resources for math retrieval, then report on the interviews of a small group of potential users to properly ascertain their needs. While preliminary, the results suggest that meta-search and resource categorization are two basic requirements for a math search engine. In addition, we implement a prototype categorization system and show that the generic features work well in identifying the math contents from the webpage but perform less well at categorizing them. We discuss our long term goals, where we plan to investigate how math expressions and text search may be best integrated.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; A.1 [**General Literature**]: Introductory and Survey; J.2 [**Computer Applications**]: Physical Science and Engineering—*Mathematics and statistics*

General Terms

Algorithm, Performance

Keywords

Math Information Retrieval, Web Classification, Niche search engines, User requirement analysis, Interaction histories

1. INTRODUCTION

While search engines help users support their general information needs, finding information for many specialized subjects and genres requires more careful attention. In this paper, we report on the user requirements study and preliminary implementation phases in creating a digital library that indexes and retrieves educational materials on math. Such a search engine would index only math materials, but draw on a diversity of materials ranging from elementary topics to current topics in mathematical research and serve an accordingly diverse range of users.

In specialized search, incorporating domain knowledge and understanding is critical in indexing, retrieving and presenting information to users. Unlike humanistic disciplines such as history and literature, mathematicians have a method of succinctly and precisely communicating findings and ideas among each other: math expressions. Math expressions – as theorems, axioms and equations – create a dual form of communication that complements the running text. To our knowledge, no studies have explored the effects of how such symbolic expressions can be exploited to address users' information needs. A motivation for exploring math search is that lessons learned can be adapted and applied to other domains that also have alternative representations than text, such as chemistry (molecular structures) and biology (DNA sequences).

The development of any digital library should be an iterative process, where cycles of gathering user requirements, designing and implementing the system, and testing are applied. In the first half of the paper (Sections 2-3), we detail our preliminary user requirements study that uncovered two clear needs for any domain-specific search: meta-search and resource categorization. A key finding of our requirements analysis is that although expression retrieval seems useful to users, simple keyword search must suffice to retrieve expressions accurately – users do not find specialized input languages usable. In the second half of the paper (Section 4), we report on the design, implementation and evaluation of the first-cut towards building a Math Information Retrieval (MIR) search engine that addresses these needs. We end by discussing the next round of development and how math expressions and text may be best handled in future work.

2. BACKGROUND

As digital libraries and resources proliferate, how scholars find, access and use information changes. Researchers, teachers, students and the general public turn to online sources for quick, indicative searches and for longer sessions of information gathering. In the current digital environment, such searches often begin as general keyword searches to large, publicly-available search engines.

However, such a search strategy works poorly for domain-specific information. Many scholarly disciplines now have a wide range of

JCDL'08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

resources on the web, in which topics can be explained at different levels: from the neophyte to the research specialist. In math, the topic of modular arithmetic serves as a case in point: simple examples can be explained to children in the guise of clock arithmetic, but specialists' needs in ring theory might also start with a similar search need. General search cannot – and probably should not – cater to the specific needs of disciplines, motivating the need for niche, domain-specific engines. Such search engines have already appeared for many media types and disciplines: for images (Flickr), patents (Google Patents), books (A9) and even math functions (Wolfram Functions Site).

Defining and characterizing such gaps between general search engines and domain-specific ones is a focus of digital library (DL) community. Such work has explored the needs of the communities of computer and information sciences, but less for other sciences and the humanities. The focus of this first half is to better define and understand this gap for the domain of mathematics. To this end, we now review how past information seeking studies inform us in the case of math, survey the major math resources and examine the current state of research in MIR.

2.1 Scholarly Information Seeking Studies

Studies of information seeking and requirements gathering are so numerous that a focused review is difficult to compile (Case's monograph [8] surveyed thousands of articles), thus we limit our review to recent studies of discipline-specific seeking.

The most closely related to our work is Brown's 1999 study on science and engineering information seeking [5]. This large-scale study surveyed faculty from several different disciplines, including math. Brown stated that mathematicians rely more heavily on monographs and older work in comparison to other disciplines. However, the study pre-dates the existence of many online interfaces to journals and databases, as well as the appearance of web-based teaching and learning resources. To our knowledge, no work since Brown's has examined math information seeking.

An alternative is to try to extrapolate results from more recent studies on other disciplines. Buchanan et al. observed the searching sessions of humanities scholars [6]. A critical finding of their work included the need for disambiguation and better refinement of domain terminology (c.f., Bates' "discipline term"), in which searches for such terms yielded thousands of hits (information overload) in the local OPAC. Wiberley and Jones [25] also observed humanists and concluded scholars (both junior and senior) "will not adopt a technology that does not promise to save time or contains no content relevant to their work". Tibbo [24], in studying historians, noted the growing influence of domain-specific websites, but acknowledged usability and accessibility problems. She recommended that such websites classify their resources and give usage instructions with their resources.

Fewer studies have connected information seeking and requirements analysis with system design. Several large scale DLs have incorporated citation linking, document chunking, authority control and discipline term / named entity linking, by both manual and semi-automated means, where these features have been stipulated by requirements analysis and/or created in response to feedback from users. Examples of such systems include Tufts' Perseus classical DL and UCSB's Alexandria georeferenced DL.

2.2 Current Math Resources

During the course of our user requirements survey (detailed later in Section 3), we collated a list of online resources that were mentioned by study participants (shown in Table 1). We characterize these math resources by type, availability, access point(s), collection scope and whether any math-specific techniques are used.

Table 1 is only indicative, as it just lists the sites reported by participants. In particular, we note that many resources on specific math topics can be drawn from smaller websites if they can be efficiently mined. Such small sites represent a majority of freely-available (i.e., subscription-free) materials largely encompassing math help and tutorial sites. A search on *modular arithmetic* in both Google and Yahoo! illustrates this point as many of the relevant resources are not part of the websites listed above.

Several aspects of this table are worth calling attention to. First, several of major databases require subscription. This hampers the accessibility for most users and hence the usability of the resource. Second, in catering for the math audience, we observe most sites do so by collecting and organizing math resources but not all of them are equally math-aware of the contents that they index. In particular, we have observed three different degrees of math-awareness:

Math-unaware: Examples include Google Books, Zentralblatt Math [17], Web of Science and MathWorld. The indexing and retrieval modules ignore the mathematical nature of their content, discarding punctuation and treating math terminology as simple tokens. For example, MathWorld can match LaTeX expressions in documents, but it does so by simple token matching, rather than recognizing LaTeX natively.

Syntactically Math-aware: Examples include Mathdex [22] and LeActivemath [21]. Such systems parse the expression to recover the syntactical structure of the math expression. Therefore, they are capable of expression matching at syntactic level and are more accurate.

Semantically Math-aware: Examples include MathWebSearch [16] and Wolfram Functions Site. Systems in this category capture not only the syntactical structures but also the semantic contents of the expressions. With this semantic knowledge, they are capable of expression manipulation to resolve the equivalence between expressions which are different in syntax but are semantically identical.

2.3 Research in MIR

Groups fielding math-specific search engines are also engaged in forward-looking research and development. From our studies of current MIR, two major areas of concern emerge: 1) how to formulate math queries, and 2) how to index and search math materials.

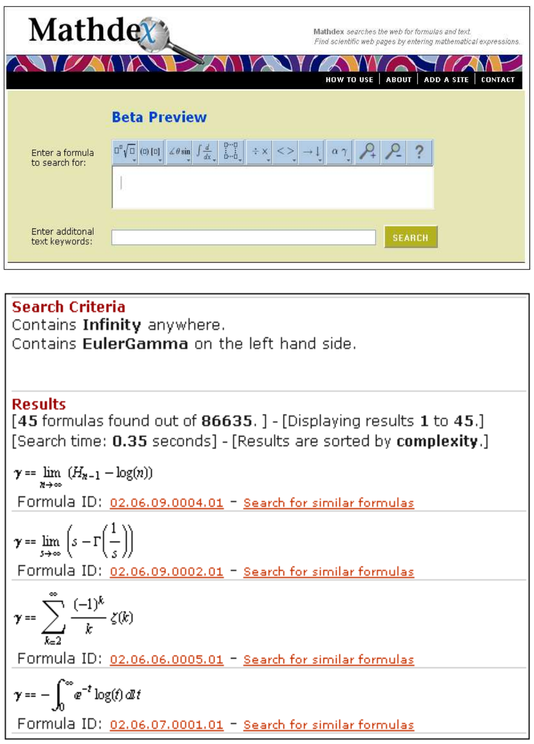
Query Language: With the keyboard serving mainly a text input device, how expressions can be efficiently entered as queries is a basic question. One straightforward way is to directly use math authoring languages like LaTeX or MathML. This method saves a lot of trouble in the process of system development since the tools for parsing expressions in such languages are readily available. In addition, it is also a favored input method for math researchers because LaTeX is the language they commonly use for paper writing. However, it still requires some work to extend such languages to cater for the specific needs of the search system. The difficulty in learning such languages may also discourage the less experienced users from using it. Currently LaTeX is used by some math-unaware search systems like MathWorld, while MathML is used in MathWebSearch, which is math-aware.

In order to enhance the accessibility of the query language to its users, math-aware search engines (e.g. Mathdex, ActiveMath and Wolfram Functions) often complement standard keyboard input with graphical, on-screen keyboards to build equations. Other approaches without GUI also exist.

Table 1: Major Web-accessible Math Resources.

Resource	Type	Availability	Access Point(s)	Scope	Math Techniques
Arxiv (*http://www.arxiv.org/*)	Publication	Open	Keyword, Metadata, Subject Classification	All	Nil
Google (*http://www.google.com/*)	Various	Open	Keyword	All	Nil
Google Books (*http://books.google.com/*)	Book	Open	Keyword, Metadata	All	Nil
LeActiveMath (*http://search.mathweb.org/*)	Various	Subscription	Keyword	Math	Expression Tree
MATHnetBase (*http://www.mathnetbase.com/*)	Publication	Open	Keyword, Subject Classification	Math	Nil
Mathdex (*http://www.mathdex.com/*)	Various	Open	Keyword, Expression	Math	Expression Indexing
MathSciNet (*http://www.ams.org/mathscinet/*)	Publication	Subscription	Keyword, Metadata	Math	Nil
MathWebSearch (*http://search.mathweb.org/*)	Various	Open	Expression	Math	Expression N-gram
Web of Science (*http://scientific.thomson.com/products/wos/*)	Paper	Subscription	Keyword, Metadata	Science	Nil
Wikipedia (*http://www.wikipedia.com/*)	Information	Open	Keyword, Subject Classification	All	Nil
Wolfram Functions Site (*http://functions.wolfram.com/*)	Information	Open	Search Criteria	Math	Function Indexing
Wolfram MathWorld (*http://mathworld.wolfram.com/*)	Information	Open	Keyword, Subject Classification	Math	Nil
Zentralblatt Math (*http://zb.msri.org/ZMATH/*)	Publication	Open	Keyword, Metadata	Math	Nil

Figure 1: User Interface of (top) Mathdex, and (bottom) results of a constrained search in Wolfram's Functions Site.



For example, [12] attempts to approximate expressions using ASCII letters while [13] examines the possibility of using a controlled set of vocabulary to write expressions in natural languages.

Indexing and Searching Techniques: The possible variations in expressing formulas and quantities give rise to difficulty in determining how to index expressions and perform matching. Even when a suitable internal representation can be given, handling search can be problematic due to variation in representation. Common approaches can be broadly classified into two groups based on whether they are text-based or not. Text-based approaches treat the math expression as text and apply standard IR techniques for both searching and indexing. Searching can be as simple as token matching (MathWorld) or pattern matching [15]. In more recent systems, Lucene, a high-performance text retrieval library, is often deployed for more sophisticated index and searching capability. For example, Mathdex stores different parts of an expressions as separate fields to allow parallel searching and flexible weighting of matches from different parts of the equation. Mathdex also ports n-gram matching techniques to math expressions search for more accurate relevance ranking.

MathWebSearch is an example of a non-text approach, where expressions are parsed into a substitution tree (more commonly used in symbolic math systems, such as theorem provers). This representation abstracts away the surface symbol and hence is able to overcome the notational variation problem which is otherwise hard to address with text-based approaches.

2.4 Unanswered Issues in MIR

It is clear from our survey of existing resources that there is a strong community interest in creating and interlinking math resources. While such materials are available, it is unclear whether the intended users are able to satisfy their math information needs using such resources. Are they adequate? Are they discoverable? It is also clear from examining current research trends that the MIR community has focused on math expression indexing and retrieval. But again it is difficult to ascertain whether such facilities are widely utilized by the community. Are such input modalities useful, or is general keyword search sufficient? Is expression matching and relevance a key factor in actual math search?

While the related information seeking literature does help us build a hypothetical profile of math seekers, it is not clear whether what the math information providers are doing actually satisfies these needs.

3. USER STUDY

To answer the questions above, we carried out our own user requirements study for MIR. Our requirements study thus had two objectives: 1) to ascertain what aspects of a math search engine are important and needed by users, as due diligence in the part of system design, and 2) to answer the questions above to find out whether the current work by the math providers really matches what math searchers need.

3.1 Study Design

While the long-term goal of our work is to build a usable math search engine with rigorous testing and a large user base, our initial user requirements study was deliberately small in scope. As such, we have chosen to use a qualitative, semi-structured interview rather than a quantitative survey instrument. We feel the interview format allows for more exploratory and productive tangential discussions to take place immediately and allows us to observe users' actual seeking process *in situ*. As such, the results we report here are necessarily preliminary and indicative, but are descriptive and allow us to posit and justify our system design. Similar study design have been used by [3], among others. Using this format, we have interviewed thirteen volunteer participants centering on students: two undergraduates (denoted as U1-2), seven graduate students (G1-7), one professor (P1) and three librarians (L1-3), all affiliated with the math department of NUS. The graduate students were recruited by a mass email and the rest were recruited by personal contact. Subjects were given a token sum for their participation. Note that the choice of population was deliberate. Our population is different from Brown's study (which surveyed only faculty) due to reasons: first we intended to extend and complement Brown's work, and second we wanted to focus on the information needs of non-experts (as Brown reported that non-search information seeking methods like personal communications play a sizable role in math research) and their providers (librarians). However, as general reference resources might cover most needs from younger students (secondary and below), we omitted them in our study.

Prior to the individual interviews, we prepared a checklist of topics (and associated probe questions) for discussion. Except for the ones on simple demographic particulars (e.g. their experience in searching for math materials), our questions loosely corresponded to the various stages of the Big6 Information Seeking Model [9]. These included what kind of materials they typically look for (Task Definition), how they approach searching (Information Seeking Strategies), what resources they use (Location and Access), as well as their expectation for a math search system (Evaluation).

We interviewed the subjects in their typical working environment so that we could observe their natural seeking behavior. After first introducing the goals of our research and disclosing the interview conditions, we conducted the interview according to our checklist. Participants were encouraged to discuss other pertinent issues and demonstrate their seeking behavior on a math topic of their choice. On average the interviews lasted 30 minutes and were not recorded; however summary notes were compiled during each interview. After each interview, we open-coded the summary notes and consolidated our findings. We continued interviewing and recruiting new participants while new findings were uncovered. Our findings stabilized after ten interviews, so we concluded the study after a final round of three more interviews.

To illustrate the richness of the interaction in our interview, we give a sample of the interaction history of a subject. Such interaction histories, together with the quotations from the subjects, served as an invaluable foundation of the analysis phase of our study.

P1, a professor, was trying to learn more about a theorem he encountered while reading a scholarly paper. He started by searching the web using the name of the theorem as separate words in a keyword search. However, part of the theorem name was a single letter which was discarded in the math search engine and the matching results were poor. When he revised the query to a phrasal search (using double quotes), the matched results were markedly better but many results only matched tangentially and were not relevant to the topic. Unsatisfied, P1 revised his query using more general terms, semantically related using his domain knowledge. The results were mixed, representing a diverse set of materials. After sorting through these for a few minutes, P1 managed to locate some relevant materials. Later, he also tried to search for relevant materials in a database of math research papers he used to visit, only to realize that it was no longer available as the library had recently canceled the subscription.

3.2 Findings

3.2.1 Information Seeking Behaviors

In our post-analysis, we organized observations according to interview topic. With regard to their own information seeking process, participants reported three main approaches for finding math materials: general keyword search, browsing math-specific resources and personal contact.

Participants noted that they searched the web using a *general search engine* querying for domain-specific math terminology (e.g., theorem or concept names such as *Helmholtz's theorem*, *differential geometry*, etc.). This approach is very popular because of its short response time and high availability, as well as the variety of information and resources it provides (as we also noted earlier in Section 2.2). On the other hand, the participants complained about its inaccuracy and lack of organization in the results. Such problems often drove participants to switch from general search engines to media-specific (Google Books) or domain-specific (MathWorld) ones. Moreover, it is often difficult to come up with appropriate search queries without deep math knowledge (*"The more you know the better [your query becomes]."*, P1). These factors often result in a time-consuming, trial-and-error process and frustration for the novice users; corroborating the phenomenon of the anomalous state of knowledge (ASK, [2]). When pressed about how organization might be improved, it was clear that standard IR topical clustering was not sought; but clustering by purpose, by resource type or by audience level.

> "It would be good if I can just see the results of a certain type." (G2)

> "Papers are often too difficult to understand." (U2)

Besides searching, participants also browsed books, journals, paper collections and encyclopedias to find relevant materials. As expected, online versions of such materials were preferred as they are more accessible. Participants felt that such secondary resources were better curated and structured, collating information from multiple sources. Participants generally searched by metadata or browsed materials classified by a standard ontology, such as the Mathematics Subject Classification (MSC). After locating possibly relevant materials, they scanned for relevant information. Participants judged this means as more rewarding although it was less accessible than search, while librarians noted that these resources are often expensive to compile, maintain or even simply subscribe to.

> "I also go to MathWorld to look for general information. It has a nice hierarchy for me to scan through." (G1)

> "Sometimes I am just too lazy to walk to the library." (G1)

> "We need to review our subscriptions to the journals and databases from time to time due to budget constraints." (L1)

Personal contact was also highly cited as a means to locate information. Students reported that they occasionally consult professors, usually as part of regular advisory meetings or as part of coursework consultation. Such sources may give explicit information or be able to refer the seeker to relevant information sources. This method was reported as highly effective but also subject to the contact's availability. It also required the student to put in effort in expressing the problem clearly, which often meant some preliminary seeking means had been tried and their utility exhausted. This finding corroborates Brown's finding that mathematicians may rely more heavily on their social network than in other disciplines.

> "I always ask my advisor in our regular research meeting. Usually he is able to tell the answer right away or give me a list of references to refer to." (G3)

These methods clearly exhibit three points along a cost/benefit curve: searching by keyword is fast but inaccurate and disorganized; browsing is comparatively easy yet less accessible and costly to compile, maintain, and subscribe to; while personal contact requires a stronger availability and query formulation commitment but is most effective. Perhaps surprisingly, participants felt that such methods acceptably satisfied their information need, but also identified the weaknesses of general keyword search as an area for improvement.

3.2.2 Mathematical Expression Input

From our discussion of current MIR research earlier, input and retrieval of math expressions is a focal point of current efforts. Although our participants expressed general interest in such facilities, when probed for specific applications, surprisingly, most could not picture a scenario where such an expression might be useful. The one potential usage was mentioned by an undergraduate was to find problem set solutions:

> "Maybe I will use it to find solutions to the problem set." (U1)

All other participants had doubts in the value of such capabilities, either due to the lack of mathematical expressions in their research domain, the inconvenience of entering expressions, or the high specificity of math expressions.

> "There are very few equations in my research domain." (P1)

> "It is rare for an important expression to be unnamed. (U2)"

> "I would prefer entering the name of the expression instead of the expression itself since it is easier." (G4)

> "I think searching with the equations is just too specific." (G5)

When asked to hypothesize about how they would prefer to input math expressions, all participants stated that they would prefer to input in LaTeX. This was tied to familiarity, as it was the math expression authoring tool of choice.

> "I think LaTeX would be a good choice since we all use it to write papers. Sometimes I even use it to communicate with my friends through MSN." (G5)

> "It would be good if I can visually preview the expression I've written." (P1)

It is worth noting that none were aware of the existence of MathML, the W3C recommendation for describing mathematics. Post-interview follow-up confirmed that this is largely due to the fact that MathML targets webpage authoring (a less familiar task) and not paper authoring (a more familiar task).

These negative findings in our survey suggests that the current MIR research focus may not really address the basic problems encountered by users of math IR, and that a cognitive gap exists between users and providers. We will return to this key point later in our discussion.

3.2.3 User Needs

What types of materials were our participants looking for? From our post-analysis, we observed that all queries involved single mathematical entities (e.g., math terminology or expression), and requirements on its content or style (i.e., format). We characterize needs into two broad categories: **information needs** which center on content (e.g., *proof of Poincare conjecture*), and **resource needs** which seek out sources in a particular format (e.g., *articles on set theory*). This distinction is similar to observations in web query analysis [4]. Table 2 gives a complete list of the identified needs.

Table 2: Types of math user needs identified.

Informational	name/alias, definition, derivation, explanation, example, problem/solution, graph/chart, algorithm, application and related entity
Resource	paper, tutorial, slides, course website, book, code, toolkit and data

By factoring together commonalities in our participants' comments, two other (usually tacit and unstated) facets of user needs also emerged in helping them to select relevant materials. *Specificity* measures how detailed the desired material is. Less specific resources are sufficient for a general, indicative understanding of the target entity while more specific ones give a thorough, informative understanding of the mathematical basis of the entity. *Experience* measures the amount of prerequisite knowledge required to understand the material. If the material is too hard for the user to understand, it is not helpful however relevant it is. These two facets are often correlated but distinct.

To understand how such needs are generated, we need to broaden our analysis to consider the context of the need, as described by the user's domain and intent.

- *Domain* refers to the (sub)discipline the user's main area of interest lies, which may be outside of mathematics. This can change the relevance of particular types of information or resources. For example, students majoring in finance may need code for simulations rather than resources describing the underlying theory; likewise, computational biologists are often interested in knowing the alias of a term in other domains.

- *Intent* refers to what the users plan to achieve with the materials. We observed the five categories of intent, each associated with a distinct usage pattern:

  Learning: Users who intend to learn generally consume both information and resources, even though their ultimate goal is the former. More importantly, such users usually have limited knowledge of the desired math entity and how to approach searching. Matching the experience level is more relevant for these intents.

  Teaching: Those teaching often already have a strong knowledge of the target math entity but require math materials such as slides or problem and solution sets to construct their own teaching resources. Materials at the right level of experience are important here too, but in the sense that they help to transmit the mathematical knowledge to target learners.

  Research: Users with an intent for research often seek primary materials as part of their literature studies and keeping updated. With a solid knowledge of the targeted entity, they employ access points such as author names and specific resource collections (e.g., Zentralblatt Math, MathWorld, Web of Science).

Collection Building: Although librarians need to collect proper materials regularly, they do not directly utilize the materials in most cases. As a result, they often need expert opinions to judge on the appropriateness of the resources collected. Despite the fact that their needs were reported as being satisfied largely by recommendation, resource categorization by type, experience and domain is likely to uncover good sources for them.

Application: This intent is often correlated with users from domains external to math. Here the user wants to find how the targeted entity can be applied or locate resources to facilitate application (e.g., toolkits, solver applet and libraries). An example highlighted to us were engineers who wished to apply results from high-level mathematics. In such scenarios, specificity is not important, but understanding how the entity can be transformed to match a concrete problem is.

3.3 Desiderata in MIR

Given the current state of MIR research and the evidence from our interviews, we feel that there is an unmet need for a math search engine. Such a system needs to address user needs more directly, catering to the intent and domain of how math materials are employed. In terms of the information seeking strategies we observed, such an engine would fill the gap between general search engines and targeted browsing of organized collections.

Will the current work in MIR be able to fill these gaps? Unfortunately, we do not believe this to be the case. As we saw, current research efforts center around expressions: their input (as queries), indexing and retrieval. From our study, it is clear that users find text input the most viable form of searching and that specialized input modalities for equations are unwieldy. According to the participants in our study, natural user-driven applications of the current MIR work may be limited, even in cases where expert users (professors and graduate students) are concerned. While it may be desirable that such an engine to be math-aware, we believe math search today has more fundamental problems that need to be addressed first.

With this in mind, we identify two immediate areas which we feel an MIR search engine should address: meta-search and resource categorization.

3.3.1 Meta-Search

Being able to search through multiple collections for materials is one of the most basic requirements for a successful math search engine. This is essential for achieving good coverage of the variety of resource types and ensuring high coverage on type-specific recall. Although there seem to be a number of different types of user needs, there are already several online collections which address certain types of needs. For example, MathWorld serves for most of the general informational needs; Zentralblatt Math for the academic articles; and other general web sites take care of some resource needs, such as tutorials, slides, and course websites. For sheer resource variety, the general web is by far the best; however, its lack of organization makes it difficult to use. It is desirable to consult specialized collections to cover materials for specific types. This is reflected in our user study as a common search pattern which a math search engine should provide support. Moreover, such specialized collections often exist with their own search engines as the sole access point with very little inter-collection linkages. Consequently users themselves have to remember the different sites for different purposes, and switch between back-and-forth when accessing them. This further adds to the burden on the users. A meta-search system addresses this by simply indexing and retrieving information across multiple collections on behalf of the users. While it is a simple requirement to fulfill, we believe such a service would be immediately beneficial to math users.

3.3.2 Resource Categorization

Our study found that the participants felt the general search engine results were disorganized and that different types of information and resources were presented together. As such, we believe a key need in math search is automatic resource categorization. A math search engine must classify materials by type automatically, ensuring that different needs requiring different types of information or resources are satisfied, without distracting irrelevant search results. From our study, we believe that orthogonal automatic text classification by specificity and (prerequisite) experience would also be helpful to narrow down relevant materials. We believe all three classifications are all feasible given the current state of the art: works have been published on genre classification (e.g., [20]) for type classification, vocabulary shift (e.g., [14]) for specificity, and reading comprehension scores for experience. Such automatic faceted classification results would need to be integrated using a suitable, faceted searching/browsing user interfaces (such as Flamenco [11]). We note that some search engines have already integrated such techniques (e.g., a search for *modular arithmetic* in Yahoo! also pulls up Yahoo! Answers content).

4. PROTOTYPE IMPLEMENTATION

Based on the user requirements and analysis, we have begun to work towards building an MIR system. From requirements interviews, the participants generally expressed that they were able to find satisfactory materials on the web, but that the mechanisms for finding or accessing them was difficult. Our plan is to index freely-available websites into a single math IR portal, centralizing access to many resources. These resources would further be categorized by resource type; that is, whether the webpage addresses an informational need or a resource one (c.f., Table 2).

To solve the meta-search criterion, we decided to take the open-source Lucene IR package as the IR framework underlying the project. The Nutch crawling package that wraps the Lucene IR library was then used to facilitate crawling steps. Sites (including those listed in Table 1) allowing spiders to index are indexed into the system. The portal itself thus provides a single point-of-access to multiple math related websites. Rather than serving any content directly, the site itself serves to drive traffic to indexed sites, only featuring a minimal amount of content on its own for its front page.

To solve the resource categorization criterion is more tricky, and is the subject of the discussion on our prototype. Manual categorization, while accurate, is labor intensive and subject to change (when the resources outdate themselves or when new materials replace or outdate old ones). As such automatic classification is preferable, and better aligned to the solution to meta-search (which is also fully automated).

As a starting point, we can send spidered webpages to a webpage classifier to categorize by resource type. The predicted resource type would be stored along with the index information and presented in the query results display to aid the user in determining relevance.

In this second half of the paper, we discuss how our system is architected in the next subsections, followed by the system's evaluation on a collected corpus of math related webpages.

4.1 Webpage Resource Segmentation

In practice however, the entire webpage is not the proper unit of granularity for math topics. During our user requirements study, we noted that many math webpages provide multiple resources. For example, a math topic page from the Wikipedia might include the topic's definition, history, proof, and applications in the real-world. While this makes the page (potentially) more useful when visited by the user, indexing is more difficult, as several resources are co-located on a single page. To deal with this problem most effectively, our system needs to first segment the individual resources on a page and then classify and index them individually.

Webpage segmentation is a problem that has many uses, and as such, also has much prior work. Simple approaches use regular expression wrappers to delimit portions of the page as a segment, but these approaches are website-specific and often break once a site is updated. As such, automated solutions have been proposed and dominate most fielded implementations. Some approaches view the web page as an XML-compliant document object model (DOM) tree (often by forcing the page through a XML validator and repairer) and determine the salience of nodes in the tree [23]. However, DOM-based models are often inflexible and sometimes reflect presentation structure rather than content structure [10]. Machine learning frameworks are also common. Works on PARCELS [19, 18] examined the use of co-training methods between style and content to better learn segment classes, and compared the differences between webpages from the same site to better differentiate main content from static site content (e.g., navigation links, site headers). For our work, we employ VIPS [7], which uses a vision-based approach independent from the DOM tree to judge coherent blocks of content. VIPS recursively divides the DOM tree of a webpage into smaller blocks using visual cues until the measured Degree of Coherence (DoC) on a block has reached a desired value (currently set to '6', as it gives the fewest segmentation errors).

4.2 Resource Categorization

Once segments are acquired, categorization is performed. We treat this as a supervised machine learning task, in which each segment is distilled to a feature vector. Manually labeled segments are then used in the training phase to generate a model, which can be harnessed to predict labels of unseen segments from new webpages.

Segments are assigned one of the ten labels as given in Table 4. Note that the first six are derived directly from our user study, while the last four are used to ensure that every segment can be labeled (coverage) while also providing additional feedback for use in classification and segmentation. We revisit this issue later in Section 4.4.

We follow the general approach in webpage classification by extracting generic features known to be successful at classifying whole webpages, and applying these to segments, in the guise of [1] which used content, hyperlink and layout features.

Table 3 gives a complete list of the features extracted. We include standard text categorization features such as n-grams, as well as some web-specific ones – features reflecting embedded images, hyperlinks as well as text formatting and layout.

Unlike webpages that have no natural sequential ordering among each other, we have observed that segments that are present within a single webpage often do follow a natural order (e.g., definitions come first, related links and pointers often come last), implicitly representing the logic of the designer. Thus we incorporate contextual features to capture ordering among segments.

To determine which features are important to the categorization, we have performed manual feature selection by adding features one by one and retaining only those which improve the performance.

Table 3: Features extracted for segment classification. * denotes selected features, see Section 4.4 for more details.

Word Features	
nGrams*	Unigram, Bigram, Trigram.
WordCount*	Number of words in the segment
Image Features	
containsImage	Whether a segment contains an image
imageFormat	The formats (jpg, bmp, png, gif, etc.) of the images in the segment.
isImage	Whether the segment is an image by itself
containsExpressionImage*	Whether this segment contains an image of a math expression
Formatting Features	
containsBoldedWord*	Whether the segment contains bolded words
containsItalicWord*	Whether the segment contains italic words
containsHeading*	Whether the segment contains headings
containsList*	Whether the segment contains lists
containsTable	Whether the segment contains tables
numberOfParagraphs*	The number of paragraphs in the segment
fontType	The names of the font types used in the segment
averageFontSize	The average font size of the text in the segment.
averageFontWeight	The average font weight of the text in the segment.
backgroundColor	The background color of the segment.
Hyperlink Features	
numberOfLinks	The number of hyperlinks in the segment
linkFileType	The types of the files (html, pdf, zip, etc.) pointed by the hyperlinks in the segment
linkTokens*	The tokens of the hyperlinks delimited by ".","\" and "&" in the segment
Layout Features	
heightToWidthRatio	The height of the segment divided by its width
sizeInPercentage	The area of the segment divided by that of the webpage
onUpperHalfOfPage*	Whether the segment is on the upper half of the webpage
onLeftHalfOfPage*	Whether the segment is on the left half of the webpage
Context Features	
nGramsFromPreviousBlock	The n-grams from the previous block
PositionInTheSequence	The number of segments before/after this one

Due to the sparsity and high dimensionality of the feature vectors generated by the n-gram features, as well as the limited number of training data we have, we use a multiclass support vector machine (SVM) learner. SVMs are well known for both their prediction accuracy and efficiency in handling such feature classes.

4.3 Corpus Development

We constructed a corpus of mathematically related web pages for development and testing of our segmentation and classification system. We first chose the scope of the corpus by selecting five common math entities: two operations ("Fourier Transform" and "Matrix Diagonalization"), two math systems ("Modular Arithmetic" and "Linear Algebra"), and a theorem ("Pythagorean Theorem"). Note that even though the aforementioned user requirements study focused only on university-level users, we chose these topic to reflect the diversity of the materials we wanted to collect in terms of type, specificity and experience. This kind of diversity is crucial to the coverage and eventual robustness of the segment categorization.

For each chosen math entity, we performed a Google web search and incorporated the first 100 results into our corpus. Out of the 500 downloaded results, 20% of them were not webpages while another 53% of them were mainly concerned with relevant resources, such as books, tools and slides, without providing any information about the math entity. We included the remaining 27%, which did contain some useful information about the math entity, to be used for annotation, development and evaluation.

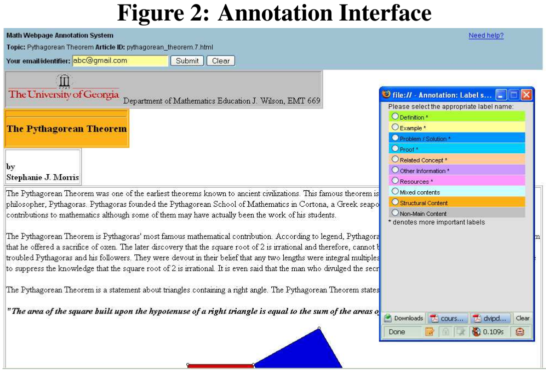
To provide ground-truth for model learning and evaluation, we then asked research group members to annotate the segments generated by passing the webpages through the VIPS segmentation system. To do this, we first developed an annotation guide that gave the definition, examples and counter-examples for each class.

Table 4: Class labels in the system (c.f., Table 2).

Name	Definition
1. Definition	A definition of the given math entity.
2. Example	An example illustrating the definition of the given math entity, how it can be applied and/or the effect of applying it.
3. Problem/Solution	A problem which requires the knowledge of the given math entity to solve and/or the corresponding solution.
4. Proof	A proof to the given math entity (usually a theorem).
5. Related Concept	Information about the concepts which are related to the given math entity.
6. Resource	Information about any other places where information/resources relevant to the given math entity can be found.
7. Other Information	Information on the target math entity that does not belong to any of the previous classes. This is to ensure that all relevant information can be labeled.
8. Structural Elements	Text or images which serve to organize or facilitate the navigation within the main content. This is to cover the structural elements in the main content like section headings, content page, etc.
9. Non-main Contents	Text or images which are not part of the main content.This is to cover anything else on the webpage like toolbars, advertisements, etc.
10. Mixed Contents	Text or images which correspond to multiple labels and shall be further segmented. This is to provide information about the segmentation errors so that measures can be taken to correct them.

Four subjects (including the first two authors) were involved in the annotation. Each subject was given a copy of the annotation guideline to read prior to starting annotation and to refer back to during the annotation process. The annotation was done through a web interface (shown in Fig.2) by clicking on each outlined webpage segment and selecting the most appropriate class from the pop-up window. In order to avoid sequential bias, the system randomizes the presentation of pages to the subjects. No specific timeframe nor time limit was set for the annotation, to ensure that subjects had ample time to do the annotation carefully. In total, 1182 annotations were done over the 135 webpages. 805 of them overlap and are used to measure inter-judge agreement.

Figure 2: Annotation Interface



We first examined whether subjects could reliably distinguish between relevant topical content and text that belonged to the whole website, or navigation. This was done by merging all annotations from label classes 1-7 into one class (relevant materials) and 8-9 as another (irrelevant) and measuring the Kappa coefficient. Kappa values range from 1.0 (complete correlation/agreement) to -1.0 (complete disagreement/negative correlation), with 0.0 indicating no correlation. The resulting average inter-judge agreement was 0.87, which indicates a high level of agreement, making this a feasible, replicable and reliable task. The Kappa coefficient remained high (0.80) when we analyzed the resulting average inter-judge agreement over all ten classes. We think this is quite satisfactory, given the fact that there are ten classes in the scheme, and we take this result as validating our coding scheme. These results show that our subjects generally agreed on whether there is relevant information in the webpage segments, although it was slightly more difficult to figure out the exact class.

Where do the confusions between subjects lie? We have observed two major confusion factors as we analyze the annotation differences. First, the annotators sometimes confuse *Definition* with *Mixed Contents* (classes 1 & 10). This happens when the definition of a math entity is given together with bits of *Other Information* such as historical background. Some then considered the segment as *Mixed Contents* while others thought the presence of *Other Information* was insignificant and labeled it as *Definition*. Second, our subjects also tend to disagree *Other Information* (class 7) with the previous six (class 1-6). We believe this acceptable due to the imprecise nature of the definition of *Other Information*.

4.4 Evaluation

We use the standard information retrieval metrics: precision, recall and F_1-measure to evaluate the classification performance and perform feature selection. To avoid overfitting, we apply 5-fold cross validation and take the average as the final result. Table 5 shows the classification performance as groups of selected features are incrementally introduced. We present our analysis of the features group by group:

Word. All the features in this group have contributed positively to the performance: the n-grams by themselves serve well as a competitive baseline for most classes, while the word count feature helps to distinguish *Structural Elements* (commonly short) and *Mixed Contents* (usually longer than normal).

Image. Different from the normal webpages, math webpages often contain a large number of math expressions as images of different types and sizes. This renders most of common image features such as containsImage, imageType and imageSize less helpful. On the other hand, however, distinguishing whether an image specifically contains math expression works very well, improving the recall significantly for *Other Information*. This is because it effectively separates the math-related content from the rest of the page.

Formatting. More often than not, *Structural Elements* are formatted differently from the normal texts on a webpage. As a result, formatting features work very well in identifying them. Nevertheless, not all the features in this group are selected. For example, font related features are found to be ineffective. This is probably due to the fact that using different font type for different content is not a common practice while changing the font size and weight can be readily done with other formatting tags.

Hyperlink. Similar to images, hyperlinks appear almost everywhere on a webpage. Therefore, the number of links in a segment is not really informative for classification. Moreover, since most of them are pointing to webpages, the types of the files pointed by the hyperlinks would not help either. Therefore, the only selected feature in this group is tokenized version of the hyperlink. This improves the classification per-

Table 5: Evaluation Results. Keys for category labels (as columns): D-Definition, E-Example, PS-Problem/Solution, P-Proof, RC-Related Concept, R-Resource, OI-Other Information, SE-Structural Elements, NC-Non-main Contents, MC-Mixed Contents. Keys for feature groups (as rows): W-Word, I-Image, F-Formatting, H-Hyperlink, L-Layout, C-Context.

Precision	D	E	PS	P	RC	R	OI	SE	NC	MC
W	.05	.38	.20	.53	.4	**1.0**	.38	.64	.66	.33
W+I	.04	.43	.33	.55	.50	**1.0**	.63	.66	.67	.34
W+I+F	.05	.46	.33	.56	.18	.67	.65	.71	.73	.36
W+I+F+H	.06	.52	.33	.58	.33	.40	**.66**	**.75**	**.75**	.38
W+I+F+H+L	.60	.55	**1.0**	**.62**	**1.0**	.40	.59	**.75**	.63	**.40**
W+I+F+H+L+C	**.80**	**.57**	**1.0**	.59	**1.0**	.40	**.59**	**.75**	.61	**.40**

Recall	D	E	PS	P	RC	R	OI	SE	NC	MC
W	**.40**	.17	**.06**	.44	.04	.06	.11	.41	.63	.52
W+I	.19	.18	.03	.44	.08	**.11**	**.60**	.41	.62	.51
W+I+F	.17	**.20**	.03	**.45**	.04	.06	.59	**.82**	.64	**.53**
W+I+F+H	.14	**.20**	.03	.36	**.10**	.08	.59	.81	.81	.52
W+I+F+H+L	.05	.17	.03	.36	.06	.06	.58	.79	**.96**	.48
W+I+F+H+L+C	.07	.18	**.06**	.36	.08	.06	.59	.77	.95	.47

$\mathbf{F_1}$	D	E	PS	P	RC	R	OI	SE	NC	MC
W	.09	.23	.06	.48	.08	**.11**	.17	.50	.65	.41
W+I	.07	.26	.06	**.49**	.08	**.11**	.60	.51	.65	.41
W+I+F	.07	.28	.06	.45	.07	.10	**.62**	.76	.68	.43
W+I+F+H	.09	**.29**	.06	.44	**.16**	.10	**.62**	**.78**	**.81**	**.44**
W+I+F+H+L	.09	.26	**.07**	.45	.12	.10	.59	.77	.76	**.44**
W+I+F+H+L+C	**.13**	.28	.06	.45	.08	.10	.59	.76	.75	.43

formance of *Related Concept* and *Non-main Contents* with respect to F_1.

Layout. Although the layout features are unable to improve the performance for F_1, they improve the precision dramatically for *Definition*, *Problem/Solution* and *Related Concepts* at the cost of recall. We think this is an advantage as precision is commonly emphasized over recall for web tasks.

Context. Unfortunately, none of the context features we have implemented are able to improve the performance significantly. We are still studying the cause and possible ways to model context appropriately.

In terms of overall performance, we can see that the current set of features is able to identify the math contents from the webpages (as indicated by the F_1 for *Structural Elements* and *Non-main Contents*) but is still very weak in categorizing them. We believe that this is mainly due to the training data and the segmentation.

The training data. There are insufficient training data for the the poorly categorized classes (F_1 < 0.4) and the distribution of positive examples for classes is skewed. Take the worst categorized class *Problem/Solution* as an example, there are only 30 positive examples in the corpus and most of them come from the same two webpages. On the contrary, for *Non-main Content*, whose best F_1 obtained is 0.81, there are close to 400 positive examples coming from practically every single page. This can be readily solved by incorporating more positive training examples from different webpages for those poorly categorized classes.

The segmentation. Webpages are often over- or under-segmented. When a page is under-segmented, it can sometimes result in the entire webpage being segmented as a single segment, and being trivially annotated as *Mixed Contents*. These errors cause noise in the training data that could be addressed with better variable level segmentation. We may be able to solve this problem if we can iteratively refine the segmentation with the labels obtained from the previous round of classification. We can merge sequences of segments which share the same label, while breaking down those labeled as *Mixed Content*.

Despite the shortcomings of this initial system, we have made solid progress in constructing the framework for an MIR system. We are currently extending our work to handle specificity and experience categorization as well. Once these aspects are finished, we will have completed a system that fulfills both desiderata, and we will be in a position to field the prototype. We plan to field it after an expanded, second round of user testing and requirements analysis, as part of our cyclical development towards creating a usable MIR system.

5. DISCUSSION

While our prototype fulfills both criteria of resource categorization and meta-search, it does not yet take much advantage of the domain of the materials: math! Earlier we asserted that MIR search engines would be more compelling if they were math-aware and could leverage this in a useful way. However, through our user requirements study, we concluded that the usability of such search methods was a problem: general users found keyword search most effective and did not feel that that inputting equations was easy.

While expert users might be satisfied with onscreen equation editors such as the ones provided in current state-of-the-art MIR engines (c.f., Figure 1), the general audience of MIR engines will not be interested in such interfaces. As the findings from our study suggest, keyword search is preferred as the access method for search due to its simplicity; however, we believe that this does not suggest expression retrieval is irrelevant. How can we make expression searching and relevance ranking relevant to users while maintaining the usability of keyword search?

We believe a method to bridge this usability gap lies in automatically correlating keywords to expressions. We propose that **Keyword-to-Expression Linking**, i.e. the resolution of expressions to terminology (e.g., $a^2 + b^2 = c^2$ to *Pythagorean theorem*) would work as a form to retrieve the dual expression form of a mathematical key phrase. Developing such a model to link keywords to expressions also helps to provide additional evidence in solving the notational variance problem which plagues the indexing of math expressions. For example, we can safely ignore the notational variance between ($a^2 + b^2 = c^2$) and ($x^2 + y^2 = z^2$), so long as they resolve to the same terminology.

This linking fits nicely with meta-search and resource categorization: the former provides abundant data for learning, while segmentation and classification results provide the information about which part of the text forms the context for a math expression and how an expression relates to a math entity. When all three are combined, we believe that math search would be improved both on the surface (better support for user pattern) and at the core (deeper understanding and better indexing).

6. CONCLUSION

In this paper, we report on our preliminary work on developing a search engine for Math Information Retrieval (MIR). In our first cycle in our iterative development, we have completed a user requirement study in MIR and identified two potential directions for future research:

Meta-Search. Future math retrieval should be able to search through isolated math collections for information and resources.

Resource Categorization. Automatic classification techniques should be employed to categorize the materials collected as to their type, specificity and prerequisite experience needed.

Between these two directions, we focus on the more difficult issue of resource categorization. Our implemented prototype uses an SVM-based classifier which extracts text, web and context features to classify segmented webpages into ten classes based on our user study. As the prototype currently yields an average F_1 of 0.36, we believe there is plenty of room for improvement. We noted the difficulties in classification were partially due to the insufficient and skewed training data as well as the segmentation errors.

We noted from our user study that math awareness would assist MIR systems in expert scenarios where notational variance and precise expression search may be needed. However, users may be unwilling to use expression input systems that are currently a focus of MIR research. We believe that a more successful approach entails building a keyword to expression linkage module that would enable expressions to be input implicitly and automatically from keyphrases.

7. REFERENCES

[1] G. Attardi, A. Gullí, and F. Sebastiani. Automatic Web page categorization by link and context analysis. In C. Hutchison and G. Lanzarone, editors, *Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pages 105–119, Varese, IT, 1999.

[2] N. J. Belkin, R. N. Oddy, and H. M. Brooks. Ask for information retrieval: Part i.: Background and theory. pages 299–304, 1997.

[3] A. P. Bishop. Digital libraries and knowledge disaggregation: the use of journal article components. In *DL '98: Proceedings of the third ACM conference on Digital Libraries*, pages 29–39, New York, NY, USA, 1998. ACM Press.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] C. M. Brown. Information seeking behavior of scientists in the electronic information age: Astronomers, chemists, mathematicians, and physicists. *Journal of the American Society for Information Science and Technology*, 50(10):929–943, 1999.

[6] G. Buchanan, S. J. Cunningham, A. Blandford, J. Rimmer, and C. Warwick. Information seeking by humanities scholars. In *ECDL*, pages 218–229, 2005.

[7] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Extracting content structure for web pages based on visual representation. In *Fifth Asia Pacific Web Conference (APWeb2003)*, 2003.

[8] D. O. Case. *Looking for Information, Second Edition: A Survey of Research on Information Seeking, Needs, and Behavior (Library and Information Science)*. Academic Press, 2006.

[9] M. B. Eisenberg and R. E. Berkowitz. *Information problem-solving: the Big Six Skills approach to library and information skills instruction*. Norwood, NJ: Albex Publishing, 1990.

[10] X.-D. Gu, J. Chen, W.-Y. Ma, and G.-L. Chen. Visual based content understanding towards web adaptation. In *AH '02: Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 164–173, London, UK, 2002. Springer-Verlag.

[11] M. Hearst. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR Workshop on Faceted Search*, 2006.

[12] P. Jipsen. Text-based input formats for mathematical formulas. In *The Evolution of Mathematical Communication in the Age of Digital Libraries, IMA "Hot Topics" Workshop, U.S.A*, 2006.

[13] F. Kamareddine, R. Lamar, M. Maarek, and J. B. Wells. Restoring natural language as a computerised mathematics input method. In *Towards Mechanized Mathematical Assistants, MKM 2007*, pages 280–295, 2007.

[14] M. Kan, J. Klavans, and K. McKeown. Linear segmentation and segment significance. 1998.

[15] M. Kohlhase and A. Franke. MBase: Representing knowledge and context for the integration of mathematical software systems. *Journal of Symbolic Computation*, 32(4):365–402, 2001.

[16] M. Kohlhase and I. Sucan. A search engine for mathematical formulae. In *Proceedings of Artificial Intelligence and Symbolic Computation, AISC'2006*, number 4120 in LNAI, pages 241–253. Springer Verlag, 2006.

[17] H. Kruger. Searching mathematics with zentralblatt math: Overview and outlook. In *Enhancing the Searching of Mathematics, IMA "Hot Topics" Workshop, U.S.A*, 2004.

[18] A. M. Lau. Advancing PARCELS: PARser for content extraction and logical structure using inter- and intra-similarity features. Technical report, National University of Singapore, 2005.

[19] C. H. Lee, M.-Y. Kan, and S. Lai. Stylistic and lexical co-training for web block classification. In *Proceedings of WIDM '04*, pages 136–143, Washington, D.C., USA, 2004. ACM Press.

[20] Y.-B. Lee and S.-H. Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *SIGIR*, pages 145–150, 2002.

[21] P. Libbrecht and E. Melis. Methods to access and retrieve mathematical content in activemath. In *ICMS*, volume 4151 of *Lecture Notes in Computer Science*, pages 331–342. Springer, 2006.

[22] R. Miner and R. Munavalli. An approach to mathematical search through query formulation and data normalization. In *Towards Mechanized Mathematical Assistants, MKM 2007*, pages 342–355, 2007.

[23] G. Newby. Information space based on HTML structure. In *The Ninth Text REtrieval Conference (TREC 9)*, pages 601–610, 2000.

[24] H. R. Tibbo. Primarily history: Historians and the search for primary source materials. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries*, pages 1–10, New York, NY, USA, 2002. ACM Press.

[25] S. Wiberley and W. G. Jones. Time and technology: A decade-long look at humanists' use of electronic information technology. *College and Research Libraries, 61, September*, pages 421–431, 2000.