You Can Lead Them to Water, But You Can't Make Them Drink: Using Crowd Sourcing to Lead Library Patrons to Extended Library Services Relevant to their Search Criteria

William B. Lund and Chad Hansen

The Lee Library at Brigham Young University has extensive resources in the forms of highly skilled subject librarians, print resources and digital collections, but getting the students to be aware of and use them effectively is a problem. To help inform library patrons of these resources librarian subject specialists have created over 150 subject guides using LibGuides[1] for subject areas, classes, and some subspecialties, which was an enormous task; but are these resources being used effectively? This paper explores and reports on a subject guide recommender system, which recommends relevant subject guides based on patron searches, using a crowd sourced folksonomy and statistical classification between subject areas and search terms. These initial results are encouraging in that 80% of the recommendations are rated as relevant by the user.

Helping Students with Research

In an age of Google and on-line searching, one of the key advantages that libraries have over on-line search engines is the wealth of experience in librarians, subject specialists, licensed digital collections not accessible through Google, and extensive print collections. In the Lee Library at Brigham Young University we have used the LibGuides[2] product from Spingshare for several years to provide an easy means for librarians to provide extended subject assistance in an on-line form as well as a way for students to quickly and easily access reference help through chat sessions, email, and phone contacts. The subject guides also include pointers to databases, websites, subject librarian contact information, and printed materials. The experience of two Canadian libraries with LibGuides can be found in (Moses, Richard 2008). As potentially useful as these subject guides are, they only provide real assistance when the patron is aware of and uses them. The number of subject guides itself (over 150) is daunting to the patron. It is frustrating to rummage through all the guides looking for the one that most directly meets the student's needs. The ideal solution is to present relevant library subject guides in conjunction with the results of a patron search; but the problem is how to take all possible search terms and associate them with the appropriate subject guides.

William B. Lund is Assistant University Librarian for Information Technology for Harold B. Lee Library at Brigham Young University, e-mail: wbl0063@byu.edu; Chad Hansen is Web Programmer for Harold B. Lee Library at Brigham Young University, e-mail: cgh@byu.edu. *The authors are appreciative to Tim Spalding and Abby Blachy of LibraryThing and to the contributors to LibraryThing for the folksonomy provided for use in this study. They also acknowledge the contributions of the Library Information Technology Division of the Harold B. Lee Library for their assistance.*

One of the frustrations of academic librarians is the students' unawareness of the vast number and size of academic research resources available to them, choosing instead to limit themselves to resources which may not be academic in nature (Griffiths and Brophy, 2005). This paper will explore using a crowd sourced folksonomy within a Naïve Bayes classifier that takes query terms and returns relevant subject guides based on association between the search terms and subject guides.

Our proposed solution is in finding associations between possible search terms and the guides through a folksonomy of terms assigned to books. Naïve Bayes classification (Lewis, 1992) allows the calculation of a degree of affinity between terms provided by the patron and subject guides through the association of metadata terms provided by patrons. Due to the ambiguity of the English language it is possible for a single term to have multiple possible associations. For example, if a user provides the search term "python" there are likely to be degrees of association with books about snakes, about a British comedy troupe, and also about a programming language, to name a few. However, if the search terms were "python programming" this would greatly reduce the association between those two terms and the books about snakes and comedy. Knowing the books which have the highest association with the search terms, the system can then identify a list of the most likely subject guides for the user.

Recommender Systems

In "Developing recommendation services for a digital library with uncertain and changing data" (Geisler, McArthur, and Giersch, 2001) state that "Digital libraries—and Web-based resource collections in general—have traditionally enabled their users to locate resources through search and browse services. Over the past decade there has been growing use of recommendation systems as a way to suggest new items of potential interest to people." Many commercial library systems today include recommender systems based on library lending patterns or through search similarities. Amazon.com and other commercial web sites have long been known for providing recommendations based on customer purchasing patterns.

(Burke 2000) describes three types of recommender systems: collaborative or social-filtering, content-based, and those based on user knowledge. Specifically, he defines content-based systems as those that "use supervised machine learning to induce a classifier that can discriminate between items likely to be of interest to the user and those likely to be uninteresting." This research will use the Naïve Bayes classifier (Lewis, 1992) , trained on the association between folksonomy terms used in LibraryThing[3] to describe books and the subjects assigned to the book through its call number.

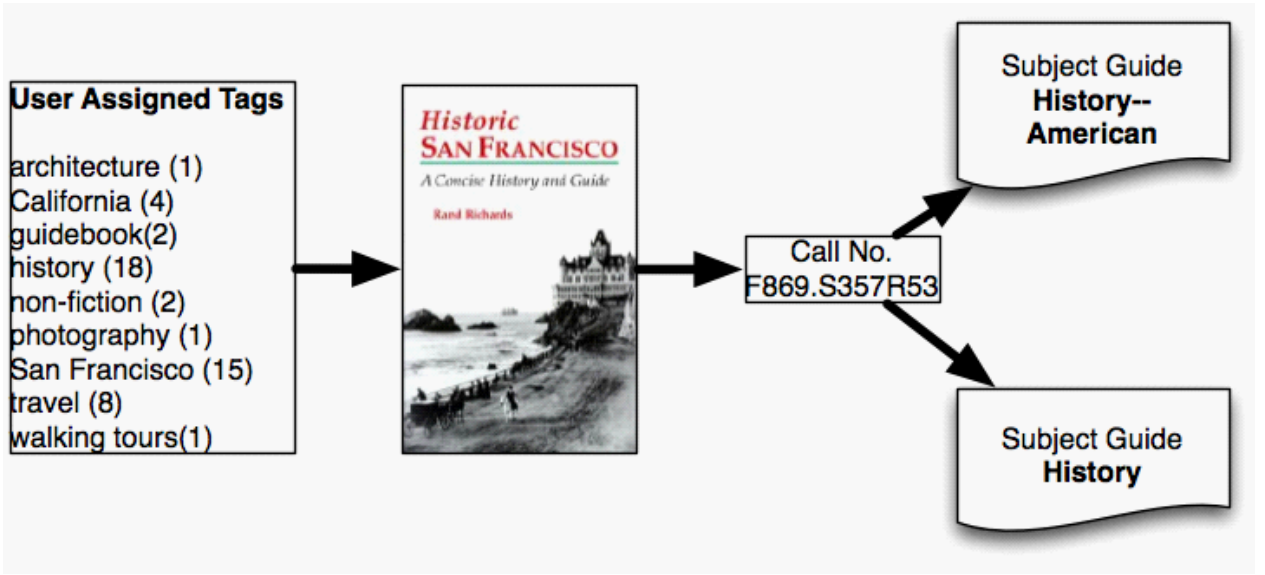
Bridging between Patron Terms and Subject Guides

Problem Context

Part of the problem in creating a recommender system for subject guides is in associating all possible terms that the patron may use in his or her search with the appropriate subject guide. It is clearly infeasible for the subject librarian to provide a comprehensive list of all possible terms that may be used to search for materials covered in a subject guide. Considering English history or biology, the number of possible terms is enormous. Another problem is that a single term, such as "python" could refer both to a programming language as well as to herpetology. Some degree of association is needed between "python" and both subjects. Further, the guides themselves are dynamic, changing as the subject librarian updates them with new resources. Using the words in the subject guides would not adequately describe the subject, as the guide is about researching a subject area, not about the subject area itself.

One possible source of these associations is a folksonomy. The power of crowd sourcing, turning a large task over to a large number of people, has been shown to be effective in many large tasks (Vukovic, 2009). This is the power behind open source projects such as PHP or Linux. LibraryThing is a community of over one million users who enter and catalog books, using descriptive subject terms that may or may not be directly related to controlled vocabularies. In previous research presented at this conference by (Lund and Washburn, 2009) it was shown that folksonomies, such as LibraryThing, use metadata terms to describe materials that are closer to what patrons generally expect than the terms in a controlled vocabulary. In short, the LibraryThing tags, are effective descriptions of the books found in those resources. In turn, books are classified with call numbers, which define the general subject area of the book. From the call number it is not a large problem to jump to a subject guide. This is demonstrated in Figure 1 for the book "Historic San Francisco: A Concise History and Guide"

FIGURE 1
Association Between User Provided Tags in Library Thing and Subject Guides Using Book Call Numbers.



by Rand Richards. In the LibraryThing folksonomy there are a number of terms assigned to this book including: architecture, California, guidebook, history, non-fiction, photography, San Francisco, travel, and walking tours, among others. On the other end of the chain, we know that the call number F869.S357R53 is found within the range associated with United States local history. In turn this subject classification is associated with two subject guides: "History—American", and "History," which are separate guides at the Lee Library. We can now say that the terms for "Historic San Francisco" have some degree of association with these two subject guides. This is represented in a database entity relationship diagram in Figure 2.

It is important to note that we are not making a strict Boolean association between these terms and the subject guides. If that were the case then every subject guide would be associated with virtually every possible folksonomy term since there is no control on how the users of LibraryThing will describe an item. Referring to Figure 3 we see the tags associated with the book "Historic San Francisco" in the form of a word cloud. Note that beside each user assigned tag is a number, which indicates the number of people who assigned that term to this work. The size of the font used for the word is based on the number of times that a term has been used to describe that work. Clearly the terms "history," "San Francisco," "travel," and "California" seem to better describe this work than the terms "architecture" or "USA." Considering the 56 million books from the LibraryThing folksonomy, each book contributes its own user supplied terms, and each term's weight as evidenced by the count of users who have chosen to use that term in their own description. Ultimately, Naïve Bayes classification, which will be described below, uses the weights of the terms from the collective books to classify user terms.

Data Preparation

One of the problems in using the folksonomy is that we may not be able to exactly match the terms used by the folksonomy contributor and the search user. Porter

FIGURE 2
Entity Relationship Diagram Showing the Relationship Between Metadata Terms, Books, Call Numbers, and Subject Guides

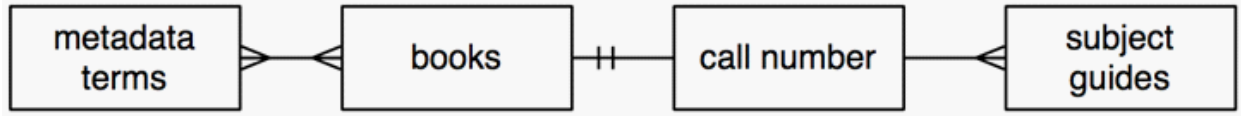


FIGURE 3
Library Thing Word Cloud Used to Describe the Work "Historic San Francisco"
(The numbers next to each tag indicate the number of people who have used that tag to describe this work)

stemming (Willett, 2006) provides a way to normalize both the folksonomy and the search terms. All words are reduced to a common root by removing prefixes and suffixes. For example, the words "helpfulness," "helpful," and "helping" are rendered by the Porter stemming algorithm as "help." This normalization preserves the initial intent of the word while providing a way to easily search across variations. All of the words in the folksonomy were processed through the Porter stemmer before being added to the database, and likewise user queries are stemmed in the same way.

To reduce the size of the database and eliminate words that do not usually relate to the meaning of a tag or query, stop words were removed both from the folksonomy tags and from user query.

Methodology to Recommend Subject Guides Based on Search Terms

Our goal is to provide, given the user's search terms, a list of subject guides closely related to those terms. This can be expressed as the probability of a subject guide (S) being associated with or given a set of search terms (T). This is written as P(S|T), however, this is difficult to calculate as it would require a priori knowledge of which subject guide the user intended when using a given set of search terms. However, relying on machine learning and using Bayes' Law we can turn it into a Naïve Bayes classification problem.

In preparation for using the Naïve Bayes classifier, consider Bayes' Law in equation 1 of Figure 4. This can be read as the probability of a subject guide (S) given a set of search terms (T) is equal to the probability of the search terms given the subject guide times the probability of the subject guide divided by the probability of the search terms. This can be simplified in equation 2 of Figure 4 by noting that we are not actually after the true probability of the subject guide given the terms, but only care about the relative weighting of the various subject guides given the terms when compared to each other. This means that we can drop

FIGURE 4
Using Bayes' Law (equation 1) and simplifying assumptions establish that the desired probability of the subject guide given the terms is proportional to the probability of the terms given the subject guide

$$(1) \qquad P(S|T) = \frac{P(T|S) * P(S)}{P(T)}$$

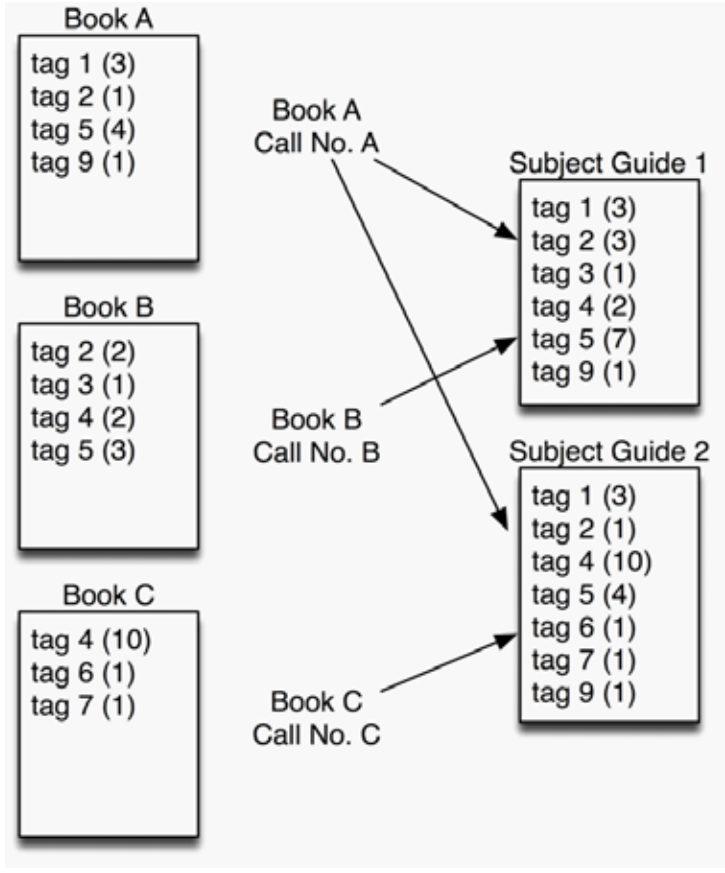
$$(2) \qquad P(S|T) \propto P(T|S) * P(S)$$

$$(3) \qquad P(S|T) \propto P(T|S)$$

the probability of the terms alone, P(T), since the probability of the terms do not change for any one query. In equation 2 of Figure 4 the symbol after P(S|T), ∝, means "is proportional to" and replaces the equals sign since we are no longer concerned with equality but rather proportionality between the two sides of the statement.

Regarding the probability of the subject guide, P(S), although we have access to the web logs for LibGuides and can use this to calculate usage of the various guides, we decided that this was biased and did not represent actually probability of a subject guide's usage if they were all equally well known. There were two reasons for this. First, students are not uniformly aware of the LibGuides which means that not all guides would have an equal chance of being used driven only by student need. And second, some LibGuides are directly referenced by faculty in their classes or in library research training, meaning that their usage may be driven by assignment rather than student need. Both of these problems would likely bias the usage numbers. Given this, we decided to give all subject guides an equal probability of use. Since this makes P(S) a constant, it can be removed from the statement as shown in equation 3 of Figure 4 since, again we are only concerned about proportionality of the subject guides.

FIGURE 5
The total usage of the user defined tags for all books within a subject guide's call number ranges are summed, giving a weight to a tag's usage within a subject guide



The Naïve Bayes classifier simply computes the probability that a subject guide would be associated with the terms. A simple implementation of the classifier would select the subject guide with the highest value of P(S|T) from all of the subject guides given the search terms. As will be explained below, it appears that a better approach is to select a number of guides with the highest probabilities.

The calculation of P(T|S), which is the probability of the terms given the subject guide, is based on the usage counts from the LibraryThing folksonomy. For each book in the folksonomy the tags and the usage counts of those tags are associated with the call number of the book itself. Each subject guide in the study has one or more call number ranges associated with it. If a book's call number falls within the range associated with a subject guide, the book's terms and their counts are associated with the subject guide. This is shown in Figure 5. The probability of a term within a subject guide, or its association with a subject guide, is the sum of usage counts of a given term divided by the total number of all terms and usage counts within the subject guide. For example, referring to Figure 5, the probability of "tag 4" in Subject Guide 1 is the total usage of tag 4, which is 2 divided by the total usage of all terms in Subject Guide 1, which is 17. So P(tag 4 | Subject Guide 1) = 2/17 = 0.12. Whereas the probability of tag 4 in Subject Guide 2 is 10/21 = 0.48. In this example, tag 4 is more closely associated with Subject Guide 2 than Subject Guide 1. This is formally represented in equation 4 of Figure 6 where $n_{t,S}$ is the sum of the usage of term t appearing in subject guide S and ||S|| is the count and usage of all terms appearing in subject guide S.

If the user provides more than one term for the search, then assuming that the terms are not identi-

FIGURE 6
Calculating the Probability of Terms Given the Subject Guide

$$P(t|S) = \frac{n_{t,S}}{||S||} \tag{4}$$

$$P(t_1, t_2|S) = P(t_1|S) * P(t_2|S) \tag{5}$$

$$P(t_1, \ldots, t_n|S) = \prod_{i=1}^{n} P(t_i|S) \tag{6}$$

$$P^*(t_1, \ldots, t_n|S) = \prod_{i=1}^{n} \left(.9\frac{n_{t_i,S}}{||S||} + .1\frac{n_{t_i,C}}{||C||} \right) \tag{7}$$

$$P^*(t_1, \ldots, t_n|S) = \prod_{i=1}^{n} \left(.9\frac{n_{t_i,S}}{||S||} + .1\frac{n_{t_i,C}}{||C||} \right) + \sum_{j=1}^{n-1} \frac{n_{t_j,t_{j+1},S}}{||S||} \tag{8}$$

cal, equation 5 in Figure 6 applies, meaning that we simply multiple the probability of the first term times the probability of the second term. This is generalized to n terms in equation 6 of Figure 6. This assumes that the query is actually of the form:

t_1, t_2, …, t_n => t_1 and t_2 and … and t_n

This presents a problem when one of the terms submitted by the user may not be present in a particular subject guide. Since equation 6 is multiplying all of the probabilities together, if any one of the probabilities is zero, which would be the case if there were no instances of that term in a subject guide, then the product of the term probabilities would also be zero. This would be the case in Figure 5 if the search query included "tag 6" which is not present in Subject Guide 1. This is not desirable, since the values of the other terms may have been quite high, relatively speaking. This problem is handled by smoothing the joint term probability using the count of the terms as they appear in the corpus of the entire folksonomy, here designated as C. The particular smoothing used here is Jelinek-Mercer smoothing as suggested in (Croft, Metler, Strohman p257. 2010). This results in equation 7 of Figure 6 in which the particular values of .9 and .1 for the subject guide count and the total corpus count are suggested by (Croft, Metler, Strohman 2010).

At this point we are considering queries as a string of single words, whereas there may be instances of user tags in LibraryThing consisting of multiple terms. For example, referring again to Figure 3, note that there are several instances of multiple word tags, such as "walking tours." Since the LibraryThing folksonomy may include multiple words in a single tag, and since we do not have an accurate way to deter-

FIGURE 7
Query Page for the Subject Guide Recommender System

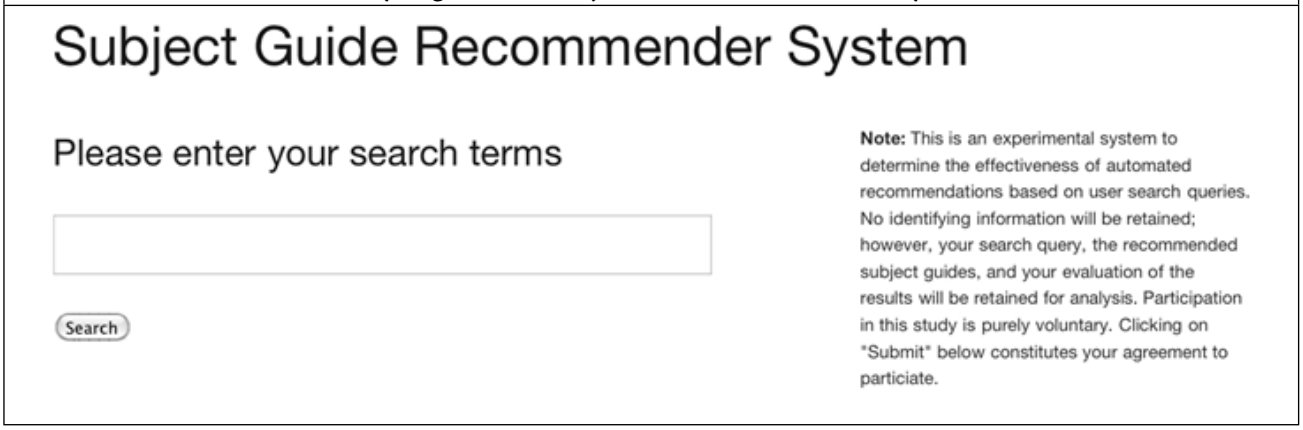
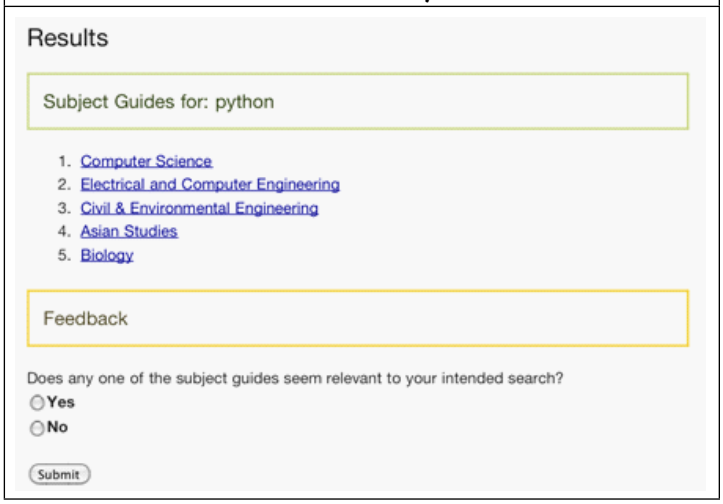


FIGURE 8
A Results Page from the Subject Guide Recommender System



mine which words of the user query may be closely associated, we calculate the probability of bigrams in the folksonomy. Our general feeling is that bigrams that match are more indicative of a close association than the single words. Consequently, we chose to "or" bigrams. This results in interpreting the user query in the following way.

$t_1, t_2, \ldots, t_n \Rightarrow t_1$ and t_2 and … t_n or $((t_1\ t_2)$ or $(t_2\ t_3)$ or … or $(t_{n-1}\ t_n))$

For example, the query "python programming language" would be interpreted as the query:

"python" and "programming" and "language" or ((python programming) or (programming language))

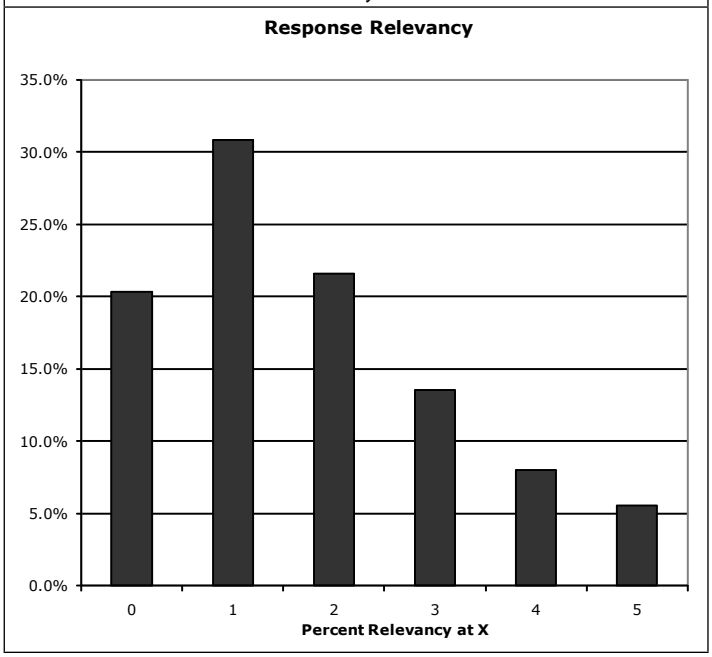
This allows us to capture bigrams which may be a stronger indicator of association with the subject guide than the individual terms alone. It is possible to extend this to trigrams and up, which is left for future work. This is captured in equation 8 of Figure 6.

Implementation and Examples

The system as described above was implemented using PHP on our library web server. Since our use of the LibraryThing folksonomy was granted only for research purposes the system has not be included in the library's main search page, but is accessed via a link indicating that this is an experimental system. The initial page is shown in Figure 7 and the response to the search on the term "python" in Figure 8. In this example the query the results returned are for the subject guides: "Computer Science," "Electrical and Computer Engineering," "Civil & Environmental Engineering," "Asian Studies," and "Biology." The system currently always returns the five subject guides with the highest value for P(S|T). It is interesting how both "Computer Science" and "Biology" are in the results set. Given the uncontrolled nature of the folksonomy, it is not unusual for there to be terms such as "Asian Studies" in the results set, which do not seem to match the search criteria at all. However, if one were to explore the folksonomy in detail, the mathematical reason for this result would be evident. This is one of the reasons why the system returns the top five results, to hopefully ensure that the relevant result appears. A more principled approach is left to further research.

As more detail is provided, the results improve. For instance, when searching on "python programming language" the results are: "Computer Science," "Electrical and Computer Engineering," "Technology & Engineering Education," "Mathematics," and "Asian Studies." In this case "Biology" was dropped from the list, most likely since the query is more precisely focused on "python" as a programming language.

FIGURE 9
The Percentage of Responses Containing Exactly X Relevant Subject Guides



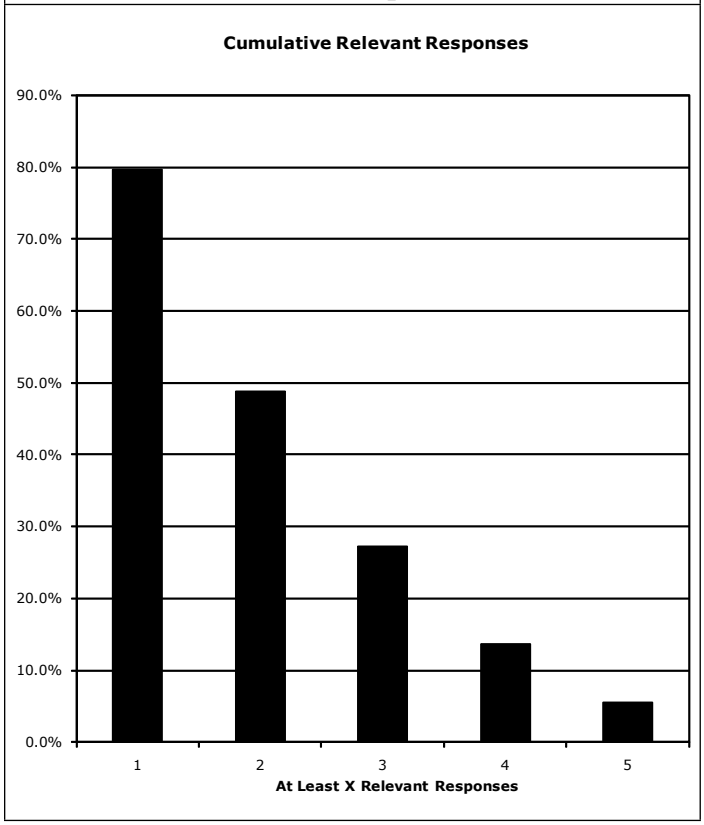
Results

As seen in Figure 8, the system collects responses from the users to determine whether the results sets are matching their expectations. The Subject Guide Recommender System web page is accessible to anyone, but we assume that the majority of the users were students and faculty affiliated with Brigham Young University. Overall 79.6% of the results sets were deemed to have at least one relevant result by the submitters of the queries with the average number of relevant results being 1.7. Figure 9 shows the distribution of the number of relevant results. Figure 10 shows the cumulative percentage of relevant responses, again reflecting that 79.6% of the responses were judged by the user to have at last 1 relevant response. Of more particular interest is the order of occurrence of the first relevant result found in Figure 11. From the result sets that had at least one relevant response, 69.8% of the time the first result was relevant. This shows that when the system is able to respond to a user's query, the first result is often relevant.

Reviewing the queries in which no relevant results were returned, ten used terms that were not found in the folksonomy at all. Those terms were: "horchata," "milage," "Mph," "guaifenesin," "warbreaker," "polenta," "cilantro," "coriander," "Lake Erie," and "Wampum." It may be a little surprising that some of these terms do not appear in the folksonomy, such as "Lake Erie" but others are very specific, such as "guaifenesin" and it is not surprising that a LibraryThing user has not used that term to describe a book.

Other failed searches simply were not closely associated with any available subject guide such as "Spruce Goose" and "Internationale Space Station" [sic]. The results set for the former query was: "Biology," "European Studies," "Juvenile Literature," "English Literature," and "History—World." Although the user was probably looking for information on the aircraft built by Hughes Aircraft at the end of World War II, the terms "spruce" and "goose" may have likely overwhelmed the results with the "Biology" subject guide coming to the top. Further, no subject guide specifically addresses aviation. The closest may be a history subject guide such as "History—World." Regarding "Internationale" although stemming is used to gather similar grammatical uses of the same word, some spelling errors would result in an unknown term. Another example of this is the query for "milage" which, according to the dictionary, is an alternate spelling to "mileage" but may be uncommon enough to be unknown in the folksonomy.

FIGURE 10
The Percentage of Results that had at Least X Relevant Responses

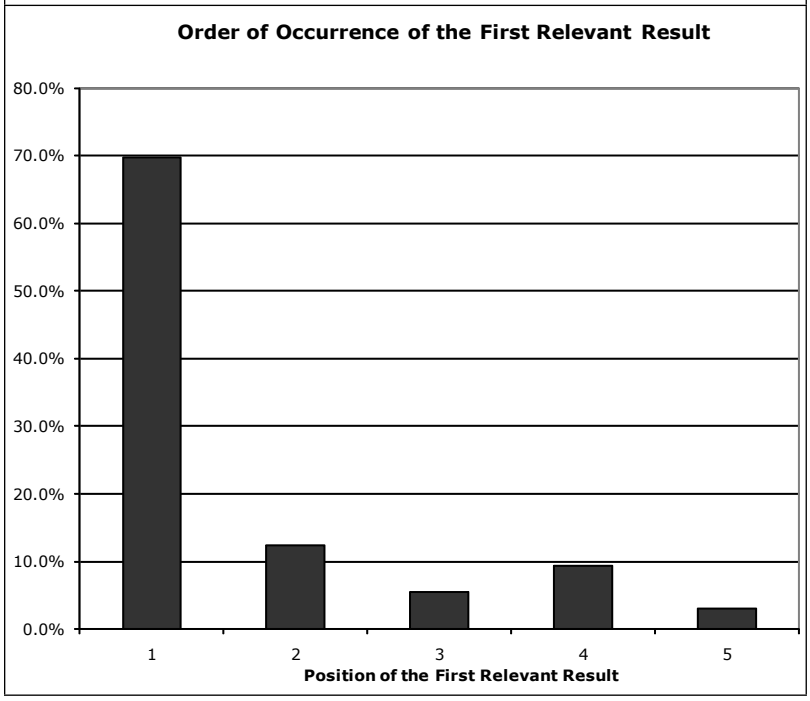


Future Research

There are several possible areas for future work. First, it would improve the results if we could identify a way to calculate the probability of a subject guide as appears in equation 2 of Figure 4. Assuming that all subject guides would be equally used is valid only in the sense that we have no information to do otherwise. Although the Naïve Bayes classifier seemed to have been effective, there are many other classification algorithms to be explored. Currently the recommender system shows the top 5 results. It is possible that more or fewer results were valid. It would be useful to find a principled approach to cutting off the list of relevant Subject Guides.

Regarding the folksonomy itself, it would be interesting to compare the search terms in our online catalog with those that appear in folksonomy itself. Is there good coverage or are there many terms from our observed searches that are missing from the folksonomy? Further, since the

FIGURE 11
The Order of Occurrence of the First Relevant Result



call number is only a single indicator of the subject of a book, would a book's controlled subject headings provide a better linkage between terms from the folksonomy to subject guides?

Most importantly, we need to find a folksonomy that we can use in a live production environment, rather than only in experimental systems. There are options available to us, some which require funding.

Notes

1. LibGuides for Libraries—Share Knowledge and Information. (2010, December 18). *LibGuides for Libraries—Share Knowledge and Information*. Retrieved from http://www.springshare.com/libguides/

2. Ibid.

3. LibraryThing. (n.d.). *LibraryThing: Catalog your books online*. Retrieved December 28, 2010, from http://www.librarything.com/

Bibliography

Burke, R. (2000). Knowledge-Based Recommender Systems. *Encyclopedia of Library and Information Systems*, *69*, 2000.

Croft, B., Metzler, D., & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice* (1st ed.). Addison Wesley.

Geisler, G., McArthur, D., & Giersch, S. (2001). Developing recommendation services for a digital library with uncertain and changing data (pp. 199–200). Presented at the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, VA.

Griffiths, J. R., & Brophy, P. (2005). Student Searching Behavior and the Web: Use of Academic Resources and Google. *Library Trends*, *53*(4), 539–554.

Lewis, D. D. (1992). *Representation and learning in information retrieval.* (PhD Dissertation). Amherst, MA.

Lund, W. B., & Washburn, A. (2009). Patrons Cataloging? The Role and Quality of Patron Tagging in Item Description. In *Proceedings of the Fourteenth National Conference of the Association of College and Research Libraries* (pp. 263–271). Seattle, Washington, USA.

Moses, D., & Richard, J. (2008). Solutions for Subject Guides. *Partnership: the Canadian Journal of Library and Information Practice and Research*, *3*(2).

Vukovic, M. (2009). Crowdsourcing for Enterprises. In *Services—I, 2009 World Conference on* (pp. 686–692). Presented at the Services—I, 2009 World Conference on. doi:10.1109/SERVICES-I.2009.56

Willett, P. (2006). The Porter stemming algorithm: then and now. *Program: Electronic Library and Information Systems*, *40*(3), 219–223. doi:10.1108/00330330610681295