ABSTRACT
This paper reports the results of a user study comparing the use of a traditional hierarchical tree interface with an associative graphical interface to select controlled vocabulary terms for document classification. The results suggest that users prefer to have both the hierarchical and associative map rather than the hierarchical tree alone and tend to ascribe more categories when using the experimental interface.

Keywords
Controlled vocabulary, digital libraries, visual interfaces, user study

INTRODUCTION
Every paper submitted to a conference or journal needs to have classifying terms assigned to it by its authors. Some, as in this conference, allow users to supply their own terms, an uncontrolled vocabulary, which often causes problems as the unrestrained use of terms results in many different terms applied to the same concept, thus making retrieval difficult. This is a familiar problem, though, and is known as the vocabulary problem (Furnas, 1987), with a user-supplied classification vocabulary being fraught with synonyms, homonyms and polysemes leading to multiple and diverse descriptions for the same resource.

Another approach that remedies the above problems is through the use of a controlled vocabulary (Rowley & Hartley, 2008). The user selects from a list or hierarchical list of controlled vocabulary terms to classify their paper. While sometimes authors often find the controlled vocabulary poorly suited to their work in newer areas and keeping the controlled vocabulary current and relevant is a significant challenge, this method makes retrieval efficient by mapping search terms to concepts.

Some organizations allow for both uncontrolled and controlled vocabulary for classification. For example, The Association for Computing Machinery (ACM) requests that when authors submit their work for consideration, they classify their work according to the ACM Computing Classification System.

That classification system is a well-planned, hierarchical controlled vocabulary (The 1998 ACM Computing Classification System, 2011). The ACM paper submission also permits users to enter their own classifiers, keywords, in addition to the reserved classification terms. This uncontrolled, flat-spaced vocabulary offers flexibility.

While the submitting authors may easily assign any uncontrolled term, the process of exploring a controlled vocabulary, particularly if it is large and complex, may overwhelm authors. A solution is needed to help authors explore a large vocabulary space and identify the best terms to describe their work. This paper explores the use of an alternative to the traditional method of presenting a hierarchical list of controlled vocabulary terms by using two different visual interfaces to help users find ACM Classification terms to classify research papers.

THE INTERFACES
The ACM classification scheme is a very large and complex vocabulary which is necessary to adequately classify papers from a variety of areas within the broad area of computing. Since it is hierarchical, it can be represented via the traditional hierarchical scheme with the higher order categories represented as folders and the terminal nodes represented as files. The use of this familiar interface does not overwhelm the user with all the terms shown at once.
and allows for the traditional drill down through the hierarchy. An example of this interface, with the classifier Hardware, Memory Structures, General (B.3.0), is shown in Figure 1.

As a supplemental aid to present classification terms, a visual map, or associative graph, was constructed to show terms that are related to the term of interest.

To determine what other terms to associate with a given term, co-occurrences of classifying terms for all the papers in the ACM digital library over a span of ten years, 1999-2009, were calculated and analyzed.

When two terms co-occur within the classification of a paper, it suggests an association between the two terms. The more they co-occur among other papers within the corpus, the stronger the association. This applies to all of the other terms that are co-cited with a term, albeit to varying degrees. Thus, if one term would be applicable for the classification of a paper, then the other terms that previous authors used may also be related, again with different strengths.

This method of unsupervised learning has been applied to co-cited authors and has been well vetted and produces meaningful results (White, 1998), and has been suggested as an interface for digital libraries (Lin, 1997). This technique using classification terms has also been suggested previously as an interface (Buzydowski, 2002).

When a user selects a term, a term seed, the system extracts all the other co-occurring terms and determines the co-occurrence of those terms. An adjacency matrix for all of the terms, with the counts of the co-occurrence of each term with each other term, is then computed. While this may result in a fully connected network, an algorithm, Pathfinder Network (Schendveldt, 1990), was used to trim the graph to its most salient links and make the visualization less cluttered.

Since it would overwhelm a user to see all of the terms that co-occur shown, only the top 20 frequently occurring other terms were used. An example of this interface, with the classifier Hardware, Memory Structures, General (B.3.0) along with the other 20 most frequently co-occurring terms and the most salient links, is shown in Figure 2.

The user interacts with the tree interface by clicking a folder to show the underlying subcategories. The user can close the category as well to remove those subcategories. The user interacts similarly with the map interface, but when the user clicks on a classifier on the left, that term is used as a term seed and a map is created on the right. When a user clicks on a node on the map, the corresponding term in the hierarchy on the left is displayed and highlighted; double clicking on a node creates a new map with that term as the seed term. The user was supplied with instructions on how to use the interface via a startup screen upon its initial use and was able to review those instructions during the process of categorizing the papers.

**RESEARCH QUESTIONS**

Given the addition of the co-occurring classifying terms represented as a map to that of the traditional hierarchal scheme, the question arises if the map interface helps users choose the correct descriptors better than the traditional, hierarchical, way.

This question is broken into three parts: 1) will users be more consistent with the document author, 2) will users be more consistent among themselves; and, 3) which interface does the user prefer?
METHODOLOGY
A comparison of two interfaces to select term(s) to classify papers using the Association for Computing Machinery (ACM) 1998 Classification System was done. The two interfaces involved the traditional hierarchical structure implemented via a file folder scheme (TREE) versus the same interface with additional graphic showing terms and their related terms derived though co-occurrence relationships (TREE+MAP).

A total of 18 subjects were contacted. Three originally agreed to be involved, but failed to complete the study for undisclosed reasons. Two partially went through the study, but had difficulty seeing the TREE+MAP interface either due to a platform issue or security settings on their machines. Consequently, a total of 13 subjects finished the study to completion, with 7 being randomly assigned the TREE interface and 6 the TREE+MAP.

The subjects were convenience sampled and needed to be involved with the computing field. Of the 13 subjects, 7 had doctoral degrees and 6 had master’s degrees. Of the 13 subjects, four had submitted papers to the ACM. The areas of the degrees were: Mathematics, Instructional Technology, Psychology, Information Science, Industrial Engineering, Computer Systems Architecture, Industrial Design, Computer Science, Business Education, Computer and Information Sciences, EE Logic Design, and Physics.

The subjects were asked to read a published paper and to categorize the paper using either the TREE or TREE+MAP interface, randomly assigned. (The author’s classifications were removed from the paper.) The subjects were then given a second paper and were shown both interfaces, which they again were asked to classify the paper. (Both papers selected were chosen with the main classifier as “general” so that no particular expertise in any area of computing would be necessary to read, understand, and classify them.) Finally, the users were given the option to choose only one of the two interfaces and were asked to indicate why they made their choice.

The measures involved with the study were the distance of the user classification to that of the original author as well as the number of categories chosen. The distance was calculated based on the number of links separating the classification tree created by a user compared to the tree generated by the author. A final measure was which interface the user preferred, along with the reason for the choice. Both the number of categories and interface preferred were simple counts. Because of the small sample size, non-parametric statistics were used to examine the differences (unless otherwise noted).

RESULTS
For the first paper, there were five classifiers used by the original author, two at level two and three at level three (A.1, H.0, D.2.6, I.6.7, K.4.0). The subjects supplied their classifications after reading the paper and the distance in terms of number of links separating them (0 = perfect classification, 28 = perfect miss) was calculated. That score was subtracted from 28 to correlate a higher score with a higher agreement. The average was 19.0 for the TREE users and 18.8 for the TREE+MAP users (p = 0.7). The sample variance between the two was 21.0 versus 17.0 for the TREE versus TREE+MAP, respectively (p = 0.6). The average number of categories chosen was 3.9 versus 5.3 (p = 0.6) and the variances were 2.8 versus 15.4 (p = 1.0) for the TREE versus the TREE+MAP, respectively. (One user selected 13 categories with the TREE+MAP interface).

For the second paper, the users were shown both interfaces and allowed to use either one to classify the paper. Six users chose the TREE and eight users chose the TREE+MAP, with two users assigned to the TREE staying with the TREE and two users assigned to the TREE+MAP staying with that interface. The second paper had only one category assigned by the author, so no measure of the distance of the author classification tree from the users’ trees was calculated. The average number of categories was 4.2 versus 6.2 (p = 0.7) and the variance was 2.2 versus 17.0 (p = 0.7) for the TREE versus the TREE+MAP, respectively. (A user selected 14 categories using the TREE+MAP interface and was not the same user who selected a large number in the first paper exercise.) Combining the two classifications yields 13 users using TREE and 13 using TREE+MAP, with the average categories chosen 4.0 versus 5.5, respectively (p=0.2; t-test).

For the final exercise, the user was told that they would have access to only one of the two interfaces for six months and they needed to choose one. Nine of the users chose the TREE+MAP and three of the users chose the TREE (p = 0.1); One user indicated that they preferred neither. In terms of the reasons, of the three that chose TREE, one indicated the map didn’t help classify papers, one said that the tree was easier to follow, and one disliked the size of the nodes of the map. Of the remaining nine who chose the TREE+MAP, the common sentiment based on their comments is that the map compliments the tree by providing additional information and that the map suggests other related terms that the user would not have thought of and thus they could select additional classifiers.

CONCLUSION
The research questions asked if the experimental interface, with co-related terms, helped a user to better classify a paper than the traditional one. While the evidence does not support a similar classification to the original, nor that one interface produced less variation, this may be an artifact of the original classification. However, the analysis does suggest that users select more categories using the experimental interface and helps users consider other categories they may not have otherwise. It could be
suggested that adding more classifiers would make a paper’s classification and eventual retrieval more efficient in terms of precision. Additionally, the users preferred the experimental interface 3 to 1 over the traditional one and this is the most significant result.

ACKNOWLEDGMENTS

This work is partially supported by NSF Award No. 0935942.

ACM graciously allowed access to the complete metadata records for the ACM Digital Library.

Since this study involved human subjects, approval was sought and granted via the IRB Board of Holy Family University.

The interfaces are available at:

http://research.ischool.drexel.edu:8080/acmsearch1/map.jsp
http://research.ischool.drexel.edu:8080/acmsearch1/tree.jsp

REFERENCES


