Latent Social Networks and Communities: An Empirical Analysis of LON-CAPA User Data

Peng Han[1], Bernd Krämer[1], and Gerd Kortemeyer[2]

[1] FernUniversität of Hagen, Hagen 54084, Germany
{han.peng, bernd.kraemer}@fernuni-hagen.de
[2] Michigan State University, MI 48823, USA
kortemey@msu.edu

Summary. the emergence and increasing popularity of social software like BBS, Blog, Wiki and Podcasting, the concept of community has shown its important role in the design and implementation of the next generation of web applications. In this paper, we report on an empirical study we performed to discover communities of practice among LON-CAPA users by analyzing log data collected over a period of three years. LON-CAPA is one of the leading online learning content management and assessment systems. We identified several kinds of social communities and showed how they may possibly influence people's selection on learning resources. Finally, we conclude our contribution and describe the future work

1 Introduction

The World Wide Web (WWW) has become one of the most comprehensive and popular forums to share knowledge and digital resources. Various online systems have been developed ranging from learning object repositories and peer-reviewed link collections like LON-CAPA [3], MERLOT [4] and Connexion[5] to research paper index services such as DBLP[6] and Cite seer[7]. While providing users with unprecedented abundant information, the ever increasing amount of resources also inevitably brings about the problem of information overload [7]. As a result, the question "How to evaluate the relevance and quality of candidate resources according to a specific search scenario?" has gained increasing attention both in academia and industry.

Keyword based search is currently the dominant approach employed by most search engines. However, as the average web users are usually not an information retrieval expert and most of them are unable to formulate boolean expressions over keywords, their information needs are often vaguely and impreciseley expressed in

[3] http://www.lon-capa.org
[4] http://www.merlot.org
[5] http://cnx.org
[6] http://dblp.uni-trier.de
[7] http://citeseer.ist.psu.edu

terms of keywords which are usually not matched well with the users' intentions. But even if a query has been well formulated, the amount of returned candidates may still be too large to be explored manually. For example, when entering the term "collaborative filtering", which refers to a small research branch of information retrieval resaerch, into Google's academic resource search engine [8], we get more than 30,000 results as of today. Beginners will probably not know where to start then.

Recently, the emergence and increasing popularity of social software like Blog, Wiki, Collaborative Tagging and Podcasting has been witnessed. They either enable users to collaboratively work on pieces of articles (such as Wikipedia[9]) or enable them to attach personalized remarks(tags)to the URL which can then be publicly accessed by other users((e.g.,Delicious[10]) for instance, supports tagging of URLs. No matter what the functionalities are, the concept of community plays a key role in the design and implementation of such systems. By grouping users with common interests or objectives together, these systems provide an effective alternative for people to share and search resources.

In this paper, we investigate the significance of the concept "community of practice" to the effectiveness of search processes for research information. We performed an empirical study to detect communities of practice among LON-CAPA users. LON-CAPA is one of the leading online learning content management and assessment system. After surveying some related research (section 2), we describe the design of our experiments and analyze the results by identifying several kinds of social communities existing among different users (section 3). Some hints on how the structure of communities may possibly influence people's selection on learning resources are then discussed (section 4). We conclude and describe future work in section 5.

2 Related Work

As revealed by related research, the inclusion of the information about the user's social environment and his position in a social network [10] of peers may lead to an improvement in search effectiveness. Collaborative filtering (CF) [2] is probably the first attempt to introduce social information into information retrieval. Its key idea is that a user is likely to prefer those items that other users with similar interests prefer. So normally a CF algorithm will first calculate the similarities between different users and then make recommendations to the target user based on the preferences of those users with high similarities.

Borrowing the idea from collaborative filtering [3, 4], we designed and implemented an experimental search engine that exploited social relationships between different users. By aggregating relevant judgments from different communities of practice, it tries to improve the search efficiency through reusing the former results in later searches using the same keywords. In order to implement this feature, users are required to join a specific community before executing a query. In [5, 6], the authors proposed a social relations mining system which combines the social networks and collaborative filtering. It focuses on extracting a social network from web pages,

[8] http://scholar.google.com
[9] http://www.wikipedia.org
[10] http://del.icio.us

finding experts for a topic and linking the searcher to the expert by a path in the social network. By extracting social links from publicly available information on the Web, it differs from other social systems by not requiring the user to sign up with a service and explicitly name his colleagues and collaborators. Similar work can also be found in [1, 9]. What all these systems have in common is that instead of directly searching for the required resources, they try first to locate those people who might have these resources.

Despite these efforts to introduce social network analysis results into information retrieval, the current approaches are far from good enough. They either require the user to maintain a separate social network dedicated to the search task or give the user too little control on how to influence the search by his social context. Similar to keywords, hyperlinks or time series, social information is just another facet [8] that the user can use to search and explore the information space. It should be easily and seamlessly combined with the other facets according to the users' requirements, so as to facilitate the whole search process. Also, the user should be able to manage and access his social information as easily as he handles other information, to define its usage and constrain the access from systems and other users.

3 Analysis of LON-CAPA Log Data

LON-CAPA is a learning content management and assessment system, originally developed by Michigan State University (MSU). Since the fall of 1992, when CAPA (a Computer-Assisted Personalized Approach) was piloted in a small physics class of 92 students, it now serves over 16,000 course enrollments per semester at MSU alone, and approximately 40,000 course enrollments system-wide, ranging from middle school to graduate level courses in over ten disciplines including astronomy, biology, business, chemistry, civil engineering and computer science.

Besides educational resources, the LON-CAPA system also keeps information about users who create, modify, assess, or use these resources. In this paper, we used a subset of the learning resources usage information kept in LON-CAPA as our experimental data. They contain 253,972 learning resources developed by 539 authors. These resources are used in 2275 courses composed by 2120 course instructors. In order to differentiate resource authors and course authors, we will call them author(s) and instructor(s), respectively in the rest of the paper. Table 1 gives a sample of the main fields in each learning resource usage record. From the record, we can see for each learning resource its author, keywords and the courses that use it.

Table 1. Sample Record of a Learning Resources Usage

Title	Authors	Keywords	Course List
Choosing Marbles Adv.	John Doe	probability	rhs_21628faa30843ceauthorl1
		repetition	uwinnipeg_134459eaa9943aeuwinnipegl1
		replacing	rhs_12090656fb7422cauthorl1

3.1 Popularity of Authors

In the first step of our statistical analysis, we found that there exist big differences among the number of learning resources provided by each author. In table 2 we list the top 10 authors who have contributed most resources in the system. For privacy reasons, we use a numerical identity to represent each author and instructor instead of their real names. From table 2 we can see that these top 10 authors have contributed more than 140,000 learning resources, which accounts for 55% of the total number of available resources.

Table 2. Top 10 Most Contributing Authors

Author ID	Number of Contributed Resources	Contribution Rank
001	89086	1
002	10831	2
003	10408	3
004	6672	4
005	4643	5
006	4167	6
007	3919	7
008	3816	8
009	3731	9
010	2943	10

In a further analysis, we studied the popularity of each author. Table 3 lists the top 10 most popular authors based on how often an author's learning resources have been used by instructors. From these results we can see that the popularity of authors does not have a direct relation with how many resources they contributed. Of the top 10 most contributing authors, only two appear in the list. To normalize the results, we define the *Normalized Contribution Popularity (NCP)* as denoted by equation 1. It represents the average usage frequencies of each resource the author contributed. The top 10 authors with the highest NCP are listed in table 4, which is almost the same as table 3 without considering the difference in rank.

Another interesting observation from table 4 is that the resources contributed by the top 10 authors with the highest NCP value account for more than 40% of the total resource usage in the system, while they add up to only approximate 6% of the whole number of available resources.

$$\mathrm{NCP} = \frac{\mathrm{Num}_{\mathrm{Used\ Resources}}}{\mathrm{Num}_{\mathrm{Authored\ Resources}}} \tag{1}$$

3.2 Author Communities

In the last section, we analyze the usage information of learning resources in LON-CAPA, which revealed the unbalance between the number of resources by an author

Table 3. Top 10 Most Popular Authors

Author ID	Contributed Courses	Popularity Rank	Contribution Rank
018	753	1	18
048	406	2	48
010	393	3	10
014	323	4	14
029	320	5	29
065	307	6	65
003	281	7	3
077	257	8	77
011	233	9	11
031	215	10	31

Table 4. Top 10 Authors with the Highest NCP

Author ID	Contributed Resources	Usage Instances	NCP	Contribution Rank
018	1930	52586	27	18
010	2943	60212	20	10
065	646	10902	16	65
077	513	7752	15	77
014	2413	25837	10	14
054	835	7726	9	54
029	1317	12266	9	29
091	404	3128	7	91
048	919	6726	7	48
011	2937	18093	6	11

and their frequencies of usage. The results indicate the existence of a small number of authors who own the majority of popular resources in the system. In this section, we look further into the data to investigate the relationship between these popular authors.

Here, we consider two authors to have some kind of relationship if their resources are used simultaneously in the same course. Or more formally, we define the Co-Contribution Association (CCA) between two users u_i, u_j by equation 2:

$$\mathrm{CCA}(u_i, u_j) = \mathrm{Count}\{C | \exists a, b \in C, a \in R_{u_i} \text{ and } b \in R_{u_j}\} \tag{2}$$

Here C is the course, while R_{u_i} and R_{u_j} are the resources contributed by user u_i and u_j, respectively. In table 5 we listed the top 20 couple of authors who have the highest CCA.

An interesting observation from table 5 is that the top 10 authors with the highest NCP also have a high Co-Contribution Association with each other. There are a total of 12 distinct authors within the top 20 couples, 9 of which are from the 10 most popular users. To describe the relationship more accurately, we define

Table 5. Top 10 Couple of Authors with the Highest CCA

CCA Rank	Author_1	Author_2	CCA	CCA Rank	Author_1	Author_2	CCA
1	018	010	274	11	077	014	135
2	018	048	261	12	029	014	130
3	018	065	225	13	077	029	128
4	018	011	182	14	018	066	121
5	048	010	173	15	014	007	121
6	018	029	170	16	029	048	119
7	029	065	154	17	048	011	111
8	065	010	146	18	031	018	108
9	029	010	138	19	065	011	107
10	065	048	137	20	010	011	105

a Strong Author Community (SAC) and Weak Author Community (WAC) with a connectivity n as equation 3 and 4 respectively:

$$\mathrm{SAC}_n = \{A | \forall u_i, u_j \in A, \mathrm{CCA}(u_i, u_j) \geq n\} \quad (3)$$

$$\mathrm{WAC}_n = \{A | \forall u_i \in A, \exists u_j \in A, u_i \neq u_j \wedge \mathrm{CCA}(u_i, u_j) \geq n\} \quad (4)$$

In table 6, we list some of the SAC and WAC we found in the LON-CAPA system, from which we can see than even for a very high connectivity such as 90, there exist such WACs or even SACs that have quite a few members.

Table 6. Sample Communities in LON-CAP

Community Type	Connectivity	Community Members
SAC	170	018, 010, 048
SAC	150	018, 029, 065
SAC	90	018, 048, 010, 029,011
WAC	200	018, 010, 048, 065
WAC	170	018, 029, 011
WAC	100	065, 011, 031, 018, 029, 048, 066

Finally, we summarize the results of our data analysis of LON-CAPA log data as follows:

Discovery 1. Most of the reusage instances in the LON-CAPA system occur for a small portion of the learning resources, which are also contributed by a small number of authors.

Discovery 2. The resources contributed by the popular authors are not only frequently used individually but also frequently used together. In other words, by their co-contribution to the same courses, popular authors form into tightly connected communities

4 Community-Aware Learning Resource Selection

Exploring the relationship between different items is an important method to help users find resources in e-commerce systems they are possibly interested in. Such relationships can be identified in different ways such as they have been frequently purchased together or have similar keywords. In existing online learning resource repositories, resources contributed by different authors are normally considered independent. However, as revealed from our findings in LON-CAPA usage data, the owner of the resources may be another useful clue for users to investigate potential interesting resources.

For example, assume that a user is searching materials for a particular course. Usually the first step is to input the keywords into the search function provided by the system to get a candidate list. After that, he or she may examine them one by one to select suitable resources. As we discussed before, one of the important issues for educational resource selection is that it is often influenced by many context factors such as learning style, teaching method and cognitive process level. This context information is often implicitly embedded in resources and not expressed literally. However, we believe that during the usage of learning resources, instructors collaboratively reveal the potential pedagogical relationship between different learning resources. For example, the resources contributed by two authors may be frequently used together because they happen to be a good complement to each other.

We believe that, by enabling the user to be aware of these kinds of Communities, we can help them to discover useful resources more easily. A typical scenario is that when an instructor selects a resource from author A, we can recommend related resources from author B whose resources are often used together with those of author A.

5 Conclusion and Future Work

In this paper, we performed an empirical study on the usage data of the LON-CAPA system, which revealed that within learning resource repositories, only a small portion of learning resources have been actively used. Moreover, these resources are contributed by a small number of authors. Furthermore, we also identified two kinds of author communities which are formed by the frequently co-used resources. Finally, we presented preliminary suggestions on how these findings may help improve the current learning resources search scheme.

Future work includes developing a prototype system to implement the Community-aware learning resource search paradigm and experiment with it in real applications.

References

1. Rodrigo B. Almeida and Virgilio A. F. A community aware search engine. In *Proceeding of 13th International Conference on World Wide Web*, pages 413–421, 2004.
2. John S. Breese, David Heckerman, and Karl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of 14th Conference on Uncertainty in Artificial Intelligence*, pages 43– 52, 1998.

3. J. Freyne and B. Smyth. An experiment in social search. In *Proceeding of 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 95–103, 2004.
4. J. Freyne, B. Smyth, M. Coyle, E. Balfe, and P. Briggs. Further experiments on collaborative ranking in community-based web search. *Artificial Intelligence Review*, 21(3-4):229–252, 2004.
5. H. Kautz, B. Selman, and M. Shah. The hidden web. *AI Magazine*, 18(2):27–36, 1997.
6. H. Kautz, B. Selman, and M. Shah. Referral web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.
7. Zengxiang Lu, Hongchao Guan, and Yanda Li. Network information filtering using bookmark service. *Journal of Software*, 11(4):545–550, 2000.
8. J. Perkioe, V. H. Tuulos, W. L. Buntine, and H. Tirri. Multi-faceted information retrieval system for large scale email archives. In *Proceeding of 2nd IEEE/WIC/ACM International Conference on Web Intelligence*, pages 557–564, 2005.
9. A. Walker, M. Recker, K. Lawless, and D. Wiley. Collaborative information filtering: A review and an educational application. *International Journal of Artificial Intelligence and Education*, 14(1):3–28, 2004.
10. S. Wasserman and K. Faust. *Social Networks Analysis: Methods and Applications*. Cambridge University Press, 1994.