How do Students Search during Class and Homework?

A query log analysis for academic purposes

Rafael López-García, Makoto P. Kato, Yoko Yamakata, Katsumi Tanaka

Department of Social Informatics. Kyoto University. Graduate School of Informatics. Yoshida Honmachi, Sakyo-ku, 606-8501, Kyoto, Japan

{rafael.lopez, kato, yamakata, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. Strong points, weak points and interests of students are precious data for their teachers, but it is hard to learn them quickly, especially when students do not cooperate in class. This paper explores a method for analysing queries of students that are allowed to search during class and homework. For this purpose, we first established six hypotheses on the queries and the expertise of the students. Then, we collected 143 queries from several lectures of an IT subject at Kyoto University. 36 students of this subject had previously been profiled before each lecture by means of questionnaires. When we checked our hypotheses against this collection of queries, we found that experts and novices often search the same way, although experts send more queries about different subjects. Some students also search contents that the teacher has not presented yet.

Keywords: query log; query log analysis; education; faculty development

1 Introduction

Passivity and lack of participation of students is one of the hardest academic challenges that teachers may find in their classes. In this case, there are few symptoms that alert the teacher about whether a certain student is having a problem or what the problem is. Furthermore, ignoring these problems or deferring their solution may have undesirable consequences such as bad academic results, loss of motivation of the members of the academic community, and reluctance to collaborate with the others.

Although teachers still have some evaluation mechanisms to tackle this problem, this unpropitious environment may also offer some other data sources that teachers hardly ever manage. Use of search engines is a good example of this, since (1) their purpose is to locate the information we are interested in, (2) it is not very difficult to collect data about their use, and (3) queries and documents are generally written as a collection of meaningful words, so it is always possible to do some kind of analysis even without employing computers.

In general, query log analysis is a difficult technique, since most of the data is anonymous, it may be incomplete and it may suffer noise. However data collected in a classroom can be easily associated to a student and combined with other sources.

In addition, by using query logs we may find other data about our students, such as the topics that motivate them, their background in other areas or just if they would be able to arouse interesting questions. This information can be used to create new conceptual relationships or to increase the participation of the students in class.

The goal of this paper is to explore a technique to analyse the queries of a classroom in order to detect and profile students that are strong in a matter, that are having problems or that are just interested in other topics. For this purpose, we apply statistical analysis to several factors in the queries we collected in several lectures of an It subject at Kyoto University, in which students have been previously profiled by means of pre-evaluation questionnaires. We also take advantage of knowing the materials used by the teachers in each moment and the contents taught in those materials, as well as the temporal relationship between queries, materials and contents.

The rest of this paper is organised as follows. Section 2 briefly discusses the related work. Section 3 studies the problem and presents our hypotheses for the analysis. Section 4 describes the educational environment in which we are collecting queries and the questionnaires we use for student pre-evaluation. Section 5 discusses the result of our analysis. Section 6 concludes and shows the future work in the matter.

2 Related Work

There is a vast literature on query log analysis [1-3], although most of these studies only address how to improve IR techniques such as re-ranking, query suggestion, etc. Some papers also try to study any social group (e.g.: children [4]) or feature (e.g.: personal interests [5-6]). However, there are not so many papers that establish a relationship between search and education [7-8], and, in fact, very few analyse a query log in order to improve education by assisting teachers or helping students.

In the context of education, most of the searches are related to a certain academic task. Therefore, analysis techniques that focus on tasks [9] may be very useful. However, most of the papers in the literature rely on concepts that are too broad. For example, they try to calculate the user intent [10-11] or even subtopics of queries [12]. Since our target is not as general as these approaches, we can use narrow these concepts and use others such as the background knowledge of a student in a certain matter or the relationship between a term and the content that is being taught.

Detecting novice and expert users is also a difficult task. Important work in the matter has been done by Lazonder et al. [13], Aula et al. [14] and White et al. [15]. An interesting feature of the latter work is that the authors try to predict expertise not only after a session, but also during it.

3 Problem and hypotheses

As we stated in the introduction of this paper, one of the essential differences between our query log and traditional ones is that it is not anonymous and we have some extra information, such as data about the students that participate in the subject (e.g.: pre-evaluation questionnaires, final scores, etc.) or the resources that are being em-

ployed in every moment. With this in mind, our analysis method is based in three assumptions regarding contents, materials and queries:

1. For any given material, we can extract the contents taught in it.

2. The materials employed in a lecture can be expressed a sequence.

3. Contents can be classified as theoretical (appear only in the slides), practical (appear only in the wiki) or both things at the same time. Queries that reference contents can be classified in the same way.

Our hypotheses about the queries sent by the students are the following:

1. Most of the queries sent by the students during the lecture will be related to the material that is being used or with any other recent one. Probability of a query to be related with a certain material will be correlated with the age of the material.

2. Queries sent by beginners will contain a lower number of teaching contents than queries sent by experts. This is based on the idea that beginners will use their searches to clarify the definition of the contents while experts will try to know more about the relationship of two or more contents.

3. If a query is strictly related to the topic that is being taught in class, queries of experts will contain a higher number of terms that are not teaching contents. This is based on the idea that beginners will just copy and paste what they do not know, while experts may add extra terms such as “definition”, “example”, etc.

4. In practical sessions, queries of beginners will focus more on theoretical content while queries of experts will focus more on practical content. This does not mean that all the queries sent by beginners will be about theoretical content, but beginners need to clarify theory more than experts before or during practice work.

5. If a query or sequence of queries is about content that is in our subject but not in the sequence of used materials (i.e: that content has not been taught yet or it is in the wiki), then the student is an expert.

6. If a query or sequence of queries is about content that is neither in the lecture nor in the materials, then the student is interested in topics that are not directly related to the lecture. This may happen because the student is trying to make a relationship between the two topics or just because of the lack of interest in the lecture.

4 Experimental setup

4.1 Syllabus and learning environment

We collected our query log in a subject at Kyoto University called “Fundamentals and Practice of Informatics A”. This subject is not about a fixed topic, but it is a collec-

tion of 15 sessions on a wide variety of disciplines whose only common point is the use of IT to solve academic problems. More concretely, the topics are:

- Document creation (Word, LaTeX).
- Web document creation (HTML, CSS, etc.).
- Cloud computing (main concepts and popular services).
- Information representation (encoding text, colours, images, mixing colours, etc.).
- Data aggregation (basic statistics and use of spreadsheets).
- Data analysis I (correlation) and II (testing hypotheses).
- Information processing I (bitmap and vector graphics) and II (natural language).
- Information retrieval.
- Database search (SQL with MySQL and phpMyAdmin).
- Data representation (XML, XPath).
- Data mining I (Introduction to the matter and basic use of R), II (association rules and clustering) and III (decision trees and introduction to Machine Learning).

However, our study does not include the sessions about "Document creation", "Web document creation" and "Information retrieval".

Note that these lectures are not completely independent, as in some cases they are continuation of the previous one (e.g.: Data analysis I & II and Data mining I, II & III).

Sessions are conducted by three different teachers, although one of them is only in charge of the two lessons in "Information processing" while another one is only responsible for the final three lessons in "Data mining".

Each lecture is 90 minutes long. It starts with a presentation by the teacher that lasts for about 35-40 minutes. After that, a practical task is presented to the students. Since the remaining time is often insufficient to complete it, students may have to finish it at home. The deadline to submit the work is often a week (in which we will continue capturing their queries). Some tasks need a certain degree of creativity (e.g. choosing a topic and creating a web page or presentation about it), while others focus almost exclusively in the technical aspect (e.g. extract the association rules from some data).

Regarding the resources, the classroom does not have a blackboard. Teachers use a wide screen in which they mostly show slides, although they may also show web pages or any other digital resource to complete their presentations. Students are provided with a printed copy of the slides and, of course, with the software they need to solve the practical work (Microsoft Office, Gimp, etc.). In addition, the subject has a website in which students can find a wiki that contains technical information to solve the practical tasks. However, teachers may demand the student to propose an example different from the one shown there (especially in lectures 13-15) and information to solve some

optional tasks may be absent. Our search engine is initially present as a sidebar of the website, although students can also access a full screen version if they want.

Students of this subject are approximately 100 freshmen (first year undergraduate) who come from very different high schools and specialities, and whose major discipline is also very disparate, from literature to computing science. Their basic knowledge in IT is also quite distinct. Only 36 of them participated in the experiment, although we have to clarify that they can skip up to 5 lectures to pass the subject.

4.2 Pre-Evaluation Questionnaires

In order to evaluate the previous knowledge of our students, we have created a questionnaire per lecture. Each questionnaire consists of several questions regarding (1) the concepts that are going to be taught in the subject, (2) the concepts that are needed to understand the ones that we are going to teach, (3) the procedures that are going to be taught in the subject (use of software, etc.) and (4) similar experience in previous subjects in high school or university.

Answers can be easily quantified (e.g.: "I have experience in doing this" / "I do not have experience in doing this" and "I can explain this concept" / "I cannot explain this concept"), and we set weights in order to balance the importance of every part, not letting concepts to be much more important than procedures and vice-versa. If the teacher wants to know some special details (for example, where the student learned a certain concept), we may include non-quantifiable questions, but the answers will not be considered in our analysis.

Students who get more than 50% in a questionnaire are considered "experts" in the matter, and we will try to distinguish their behaviour from the rest of the students, to whom we will call the "novice". According to this system, "experts" change in every lecture. Fig. 1 shows the distribution of expert students and novice students:

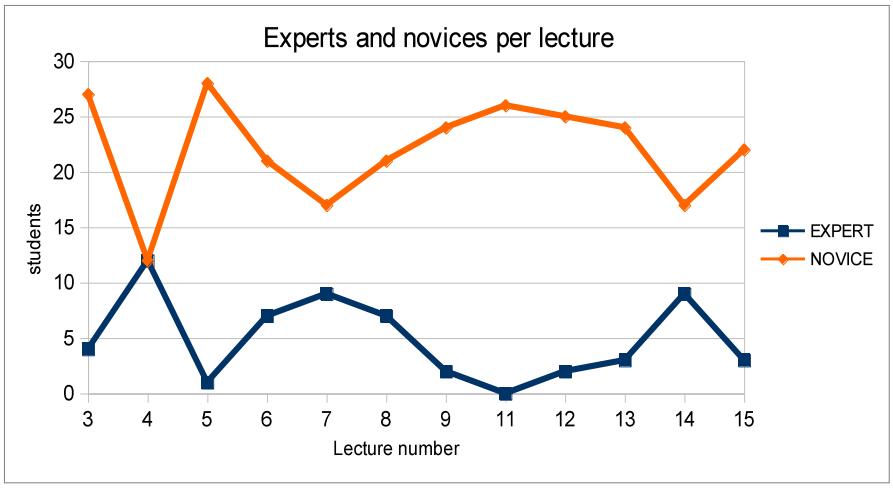

Fig. 1. Distribution of expert students and novice students per lecture

5 Analysis

We collected 181 queries in 12 lectures. However, some of them are just a visit to the previous or next page of results of a previous query, or a re-execution of the same query. If we count all these navigations as a single query, our query log is reduced to 143 queries. From these, only 30 (20.97%) were submitted by expert students.

The first impressing result we have obtained is that, in general, only 77 queries (53.84%) are related to the topic of the lecture. Expert students tend to send more queries that are not related to the lecture, and this result is statistically significant (p-value of 0.0112, 0.0198 with Yates correction). Table 1 shows this distribution:

Table 1. Are queries of students related to the lecture?

	Expert	Novice
Not related to the lecture	20 (13.99%)	46 (32.17%)
Related to the lecture	10 (6.99%)	67 (46.85%)

Contrary to what we though when we formulated hypothesis 1, from the 47 queries sent during the presentations of the teachers, only 30 (63.83%) are related to the topic of the lecture and only 23 (48.94%) are related to the slides. In fact, it happens that in the last lecture we received a sequence of 5 queries which are related to the contents of the wiki, but not to the contents shown in the slides. There is no significant difference between novices and experts (p-value of 0.2871, 0.5421 with Yates correction).

In any case, we analysed if the queries that are related to the lecture are synchronised with the slides, obtaining the result shown in Fig. 2. Column "delay" represents the difference between the current slide and the slide that is related to the query of the student. A value of 0 means they are the same slide, while a negative value means the slide related to the query has not been shown yet.

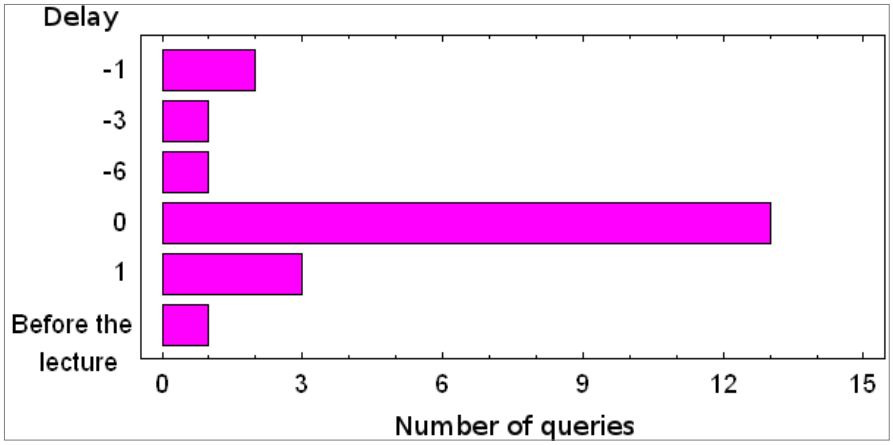

Fig. 2. Synchronization between queries and the current slide

The first result we can observe in this figure is that query sessions are normally very short, as students abandon the search if the teacher changes the slide.

The second result we find is that there are students that send queries about topics that the teacher has not taught yet. However, there is not significant difference in their expertise (p-value of 0.3679), and, in fact, there are only two queries per class of student (one expert anticipates 1 slide, another one anticipates 6 slides, one novice anticipates 1 slide and another one anticipates 3 slides). In addition, a novice anticipated a query before the lecture started. Therefore, we cannot confirm hypothesis 5.

Regarding the number of words that are teaching content ("content terms") in the queries that are related to the lecture, there is no statistical significance between experts and novices (p-value of 0.5638). However, the number of queries of experts may be too low to say something about hypothesis 2. Table 2 shows the distribution of queries according to content terms and expertise.

Table 2. Number of content terms vs. expertise of the student

	Expert	Novice
0 content terms	2 (2.60%)	6 (7.79%)
1 content term	5 (6.49%)	45 (58.44%)
2 content terms	3 (3.90%)	14 (18.18%)
3 content terms	0 (0%)	2 (2.60%)

With regard to the number of words that are not teaching content ("no content terms") in the queries that have at least one content term, there is not significant difference between experts and novices either (p-value of 0.2979). However, as in the previous case, the number of queries of experts may be too low too. Therefore, we cannot confirm or discard hypothesis 3 either. Table 3 shows the distribution of queries according to the no content terms and the expertise of the students.

Table 3. Number of no content terms vs. expertise of the students

	Expert	Novice
0 no content terms	4 (5.56%)	40 (55.56%)
1 no content term	4 (5.56%)	14 (19.44%)
2 no content terms	0 (0%)	9 (12.50%)
3 no content terms	0 (0%)	1 (1.39%)

Hypothesis 4 is related to the theoretical or practical orientation of the queries. To verify this hypothesis, we divided the queries in 4 categories: theoretical (the content appears only in the slides), practical (the content appears only in the wiki), both (the content appears both in the slides or the wiki) or none (the query is not related to the lecture). Once we excluded all the queries that are not related to the lecture, we could not find any statistical significance between these categories (p-value of 0.4991), but, once again, we have to make clear that the number of queries of experts is too low to discard the hypothesis. Table 4 shows the distribution of the queries according to their theoretical or practical orientation and the expertise of the students.

Table 4. Theoretical or practical orientation of quereis vs. expertise of the students

	Expert	Novice
Theoretical	1 (1.35%)	2 (2.70%)
Practical	3 (4.05%)	27 (36.49%)
Both	5 (6.76%)	36 (48.65%)

We have also checked the practical queries that were sent during the presentations of the teachers and they were always sent by novice students. Therefore, we can strongly reject hypothesis 5.

Queries that are not related to the lecture seem to provide an inestimable source of data about the interests of the students, supporting hypothesis 6. For example, we have located a session of one student with approximately 50 queries (including navigations between result pages and image search) containing the terms "Art Noveau", "Art Deco", "Alfons Mucha" and "William Morris". After analysing the homework of the student for that lecture, we found that it was a presentation about the aforementioned artistic movements and their most important representatives, providing us with the evidence of the interest of the student in that topic.

Another example is a student who queried "concept diagram" (5 queries related to that topic) and "Lifestyle diseases" (one query) and whose homework was a conceptual map about lifestyle diseases.

In addition, there are students that searched terms such as "Pikachu" (a character of the Japanese animation "Pokemon") or "sleep", but we could not find any evidence on their homework that confirms their interest in the matter. In the case of the first student, the task of that lecture was editing an image, so it is possible that the student considered that topic but discarded it because of the ban that the teachers established on copyrighted images.

Apart from the studies whose objective is to confirm or discard our hypothesis, we have also checked some other factors such as: (1) the type of characters used by the student (Japanese, Western or mix), (2) the use of the wiki, and (3) the number of queries that ended with a result being clicked by the student. Although in principle we detected significant difference in the two first tests, these differences were due to the fact that expert students tend to send more queries that are not related to the lecture. As soon as we analyse only the queries that are related to the lecture, the difference disappears. The third test also did not have any positive result.

6 Conclusions and future work

This paper has explored a statistical query log analysis for detecting the strong points, weak points and interests of students, especially of those who are often silent in class. The main differences between our methods and traditional ones are that (1) we employ not only queries, but also the materials and contents the teacher uses in class and (2) we consider the temporal relationship between the queries and the materials that are used in the lecture.

Our system makes some assumptions about the contents and the materials. For example, (1) we know what materials the teacher has been using in every moment, (2) we can express these materials as a sequence and (3) we can classify the content into theoretical, practical or both things at the same time.

With the aforementioned assumptions in mind, our paper presents a list of hypotheses about the queries of our students. These hypotheses consider the difference in behaviour we expected between expert students and novice students. For example:

- Queries of students are related to and synchronized with the slide that the teacher is currently showing in the presentation.
- Expert students try to establish more relationships between teaching contents.
- Expert students add terms that are not teaching contents to their queries in order to find what they expect with a higher probability.
- Queries of novice students will contain more theoretical contents as they will need to review them in order to solve their practical work.
- Expert students try to anticipate the contents that the teacher was going to show.
- Queries that are not related to the lecture can show the interest of the student in other matter.

However, in the experiment we performed in an IT subject of Kyoto University, we found that there is not a significant difference between the experts and the novices in most of these matters. In fact, the number of queries sent by experts is often not enough to offer any conclusion in the matter.

Regarding the temporal relationship between queries and materials, we can say that, when queries are related to the lecture, they often are related to the slide that the teacher is showing in that moment. Search sessions of the students are short. However, there are students that search for terms that are related to the lecture but the teacher has not presented yet. With this result we can set two future goals: (1) find a search interface that makes querying easier and (2) try to advance materials for students that are faster than the class.

Another significant result we have found is that expert students tend to send more queries that are not strictly related to the lecture. With this result, a teacher can try to use this kind of queries to make their authors participate more in the class, just by establishing relationships between the topic of the lecture and the topic of the queries.

Regarding the hypotheses that we could not confirm, we can conclude that, in such reduced environment, there is not statistical significance in the number of terms that are teaching contents or that are not teaching contents. However, in our future work we will add a new hypothesis about the quality of the contents: do experts use concepts that are more specific than the ones used by the novice students?

Another work is to analyse the navigations the students did and the documents they visited.

Acknowledgment

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013 and 23300311) from MEXT of Japan.

We would also like to thank Professors Masatoshi Yoshikawa, Yasuhito Asano and Masayuki Murakami for their inestimable support to our research.

References

1. White, R.W., Huang, J.: Assessing the scenic route: measuring the value of search trails in web logs. In: Proceedings of SIGIR '10, pp. 587-594. (2010).
2. Liu, Y., Song, R., Chen, Y., Nie, J.Y., Wen, J.R.: Adaptive Query Suggestion for Difficult Queries. In: Proceedings of SIGIR '12, pp. 15-24. (2012).
3. Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y.: Probabilistic Query Expansion Using Query Logs. In: Proceedings of the 11th International Conference on World Wide Web. (2002).
4. Duarte Torres, S., Hiemstra, D., Serdyukov, P.: Query log analysis in the context of information retrieval for children. In: 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. (2010).
5. Pu, H.T., Chuang, S.L., Yang, C.: Subject categorization of query terms for exploring Web users' search interests. In: Journal of the American Society for Information Science and Technology, vol. 53, no. 8, pp. 617–630. (2002).
6. Limam, L., Coquil, D., Kosch, H., Brunie, L.: Extracting User Interests from Search Query Logs: A Clustering Approach. In: Proceedings of DEXA Workshops 2010, pp. 5-9. (2010).
7. Liu. H.: Learning by Searching. In: K. McFerrin et al. (Eds.), Proceedings of Society for Information Technology & Teacher Education International Conference 2008, pp. 3843-3844. Chesapeake, VA: AACE. (2008)
8. Howard, P., Massanari, A.: Learning to search and searching to learn: Income, education, and experience online. In: Journal of Computer-Mediated Communication, vol. 12, no. 3, article 5. (2007).
9. Buscher, G., White, R.W., Dumais, S.T., Huang, J.: Large-scale analysis of individual and task differences in search result page examination strategies. In: Proceedings of WSDM '12, pp. 373-382. (2012).
10. Broder, A.: A taxonomy of web search. In: ACM SIGIR forum, vol. 36, no. 2, pp. 3-10. (2002).
11. Rose, D.E., Levinson. D.: Understanding User Goals in Web Search. In: Proceedings of the World Wide Web Conference (WWW 2004), pp. 13-19. (2004).
12. Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang Q., Orii, N.: Overview of the NTCIR-9 Intent task. In: Proceedings of NTCIR. (2011).
13. Lazonder, A.W., Biemans, H., Wopereis, I.: Difference between novice and experienced users in searching information on the World Wide Web. In: Journal of the American Society for Information Science, vol. 51, no. 6, pp. 576-581. (2000).
14. Aula, A., Jhaveri, N., Käki, M.: Information search and re-access strategies of experienced web users. In: Proceedings of WWW, pp. 583–592. (2005).
15. White, R.W., Dumais, S.T., Teevan, J.: Characterizing the Influence of Domain Expertise on Web Search Behavior. In: Proceedings of WSDM '09, pp. 373-382. (2009).