Academic search in response to major scientific events

Li, X.; de Rijke, M.

[Link to publication](#)

Academic Search in Response to Major Scientific Events

Xinyi Li and Maarten de Rijke

University of Amsterdam, Amsterdam, The Netherlands
{x.li,derijke}@uva.nl

Abstract. In this paper, we look at the search behavior of users of an academic search engine and in particular, their query patterns following the occurrence of major scientific events. We select Nobel Prize announcements as major scientific events and observe how academic searchers behave in response to their occurrences. We have found unique trends for the academic searchers, which are different from users of a web search engine. Moreover, academic searchers have similar query trends even for different topics, showing a commonality in their query behavior. The insights gained in this paper highlight the unique behavior of academic searchers and the differences from users of general web search engines. The findings may have implications for our understanding of the search behavior of academic searchers as well as for search engine design, for instance, concerning approaches to address information needs around major scientific events.

Keywords: Academic search; Query trend; Web search

1 Introduction

Academic search engines are the most commonly used tools to acquire research information [5]. There have been a series of studies on users of academic search engines, including surveys or small scale user studies [12–14]. However, very few studies look at academic searchers by studying a real transaction log of an academic search engine. In this paper, we look at the transaction log of a major academic search engine and focus on one aspect of academic searchers: their query patterns in response to the occurrence of a major scientific event.

More specifically, we provide answers to the following research questions: (1) Upon a major scientific event, what are the query patterns of academic searchers? (2) How is the trend in academic search compared to users on a web search engine? (3) How do the query trends of academic searchers with different topical interests compare?

We take the Nobel Prize announcements as typical cases of major scientific events. To observe the query trend of users, we first select query candidates that represent the events by a two-step approach. The approach utilizes co-occurrences of queries and requires no domain expert.

For these queries, we observe their frequencies upon the prize announcements. We use transaction log data from ScienceDirect[1] to study academic searchers and statistics

---

[1] http://www.sciencedirect.com/

from Google Trend[2] as our source of web search query frequencies. Obviously, users of a web search engine may include some academic searchers. Therefore the behavior that we observe on a web search engine is a combination of both ordinary users and academic searchers.

We conduct a series of trend analyses and cross correlation studies to answer the research questions. We find that academic searchers have a unique query pattern in response to a major scientific event, which shows a gradual steady rise. This trend is different from users of a web search engine; it exhibits a spike of surging interest in response to the event's occurrence and then declines over time. Correspondingly, we find that it is not possible to predict query trends in academic search from those in web search on the same topic, and vice versa, due to their distinct user behavior. However, users within academic search have similar query patterns even on different topics. This shows a certain degree of behavioral commonality among academic searchers.

The findings in this paper reveal query trends of academic searchers; query trends of academic searchers are different from those of users of a web search engine. To the best of our knowledge, this is the first paper of its kind that examines the querying behavior of academic searchers on major scientific events by a query log analysis. The findings in this paper shed light on the search behavior of academic searchers, and may help search engine designers produce better document rankings that more closely correspond to their users' interests.

2 Dataset

We study a transaction log from the ScienceDirect search engine, which primarily covers physical sciences, engineering and life sciences. Collected from September 28, 2014 to March 5, 2015, the query log contains more than 39 million records of traffic via institution-authorized access as well as personal access. The former refers to users in a certain IP range (e.g., from a research institution) denoted as IP-users, and the latter refers to users that log in or access the search engine from outside the institution, i.e., non-IP users. Two thirds of the query traffic comes from IP-users. The data statistics of our log are shown in Table 1.

Table 1: Query length statistics in word count.

	#N	min	max	mean	median
Sciencedirect	39M	1	419	3.77	3

3 Related work

Academic search deals with the indexing and retrieval of information objects (papers, journals, authors, . . .) in the domain of academic research. Prior to the advent of the

---

[2] https://www.google.com/trends/

world wide web, academic search engines were mostly restricted to library usage and, often, only pre-programmed searches were supported instead of online queries. A typical example is MEDLINE [11], which debuted in 1971. With the rising popularity of the web in the 1990s, online academic search engines were developed and have been serving a larger user base. Several earlier studies have looked at behavior of researchers on modern academic search engines [12–14], either via surveys or user studies, and they are restricted to a small sample of users. There are a few log analyses on search engines of digital libraries [4, 6, 7], but they focus on basic usage statistics and lack insights on user behavior. Recently through large-scale log analyses, failure phenomena have been investigated in academic search queries [10]; another recent study reveals correlations between query reformulations and topic shift [9].

Concerning the search behavior of academic searchers, little is known about query trends in response to a major scientific event, and how those trends differ from users of general web search engines. This paper differs from previous work in academic search by conducting a log analysis to find out query trends of academic searchers in response to major scientific events and comparing those trends against those of users of a web search search engine.

4 Approach

In this section we describe the approaches we take to answer the research questions. We choose three Nobel Prize announcements to represent major scientific events. We select topics that are covered in the academic search engine's database: chemistry, physics and physiology/medicine. To study academic searchers' query patterns, we first select query candidates that represent these topics in a 2-step approach. Then for each topic, we observe how the query frequencies change over time. Through time-series analysis, we compare the trends and correlations between different topics, and also compare academic search against web search.

4.1 Query candidate generation

We extract bi-grams and tri-grams from the news announcement excerpts on the Nobel Prize website[3] as potential query term seed set. For each prize/topic, we search the bi-grams and tri-grams through Google Trend and exclude those that are either irrelevant to the Nobel Prize topic (no mention of "Nobel Prize" in the top 10 results of the Google Trend "Related Topics" panel) or too infrequent in the search log (as is indicated in the results of Google Trend). Here we show a positive example of a relevant query: when entering one of the tri-grams in the chemistry topic term seed set, Figure 1 shows that the query is related to the topic Nobel Prize, therefore it is deemed as relevant. Meanwhile we also extract the names of the award recipients and treat them as relevant query candidates. After this first step we have obtained some relevant query candidates.

---

[3] See `https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2014/`, `https://www.nobelprize.org/nobel_prizes/physics/laureates/2014/`, and `https://www.nobelprize.org/nobel_prizes/medicine/laureates/2014/`

Table 2: Query candidates for each topic.

Category	Theme queries
Chemistry	"super-resolved fluorescence microscopy", "fluorescence microscopy", "super-resolution microscopy", "confocal microscopy", "Eric Betzig", "Stefan W. Hell" and "William E. Moerner", "fluorescence microscopy principle", "dark field microscopy"
Physics	"blue light-emitting diodes", "light-emitting diodes", "Isamu Akasaki", "Hiroshi Amano" and "Shuji Nakamura"
Physiology/medicine	"brain positioning system", "brain positioning", "John O'Keefe", "May-Britt Moser", "Edvard I. Moser"

In the second step we expand these query candidates to a larger query set. We search these relevant query candidates by using Google Trend, and use its "related queries function" that shows queries being searched at the same period. To avoid false positives, we only perform one round of expansion, and choose a relatively high threshold of 150% (meaning search frequency increase by at least 150%) of relevant queries. Of these expanded queries, we remove the queries that are either generic or navigational such as "pubmed", "nobel prize 2014."

In this way we have collected the relevant query candidates related to each topic, as shown in Table 2.

Fig. 1: Related Topic Panel on Google Trend, for the query "super-resolved fluorescence microscopy."

For academic search, we examine queries in the ScienceDirect log that contain the terms of each query candidate and obtain their query frequencies over time.

4.2 Query trend of academic searchers

To answer the first research question, we look at the query trends of academic searchers. We analyze how trends change after a Nobel prize announcements. The timespan we study starts from one week prior to the prize announcements, and ends at the last week of the dataset.

4.3 Difference between academic searchers and users of a web search engine

We take Google Trend as a proxy to observing users on a web search engine. We compare Google Trend statistics with the academic search engine by aligning them to the same timespan starting from one week before the prize announcement. When there are many relevant queries to one topic, it is difficult to leverage the statistics of them to represent that topic, since we do not have the query counts from Google. In this scenario we utilize the "relevant topic" statistics only, which is a function of Google Trend that summarizes the trend of the relevant queries for that topic. For chemistry we use the topic "Super-resolution microscopy." For physics, which has only a few queries, we use the representative terms "blue light-emitting diodes" due to its high correlation with the Nobel Prize topic.

4.4 Query patterns of academic searchers and users in web search in different topics

In this part of the analysis we aim to uncover whether there is a correlation between their query patterns in different topics. The hypothesis is that there may be some behavioral commonalities within the academic searcher community, so their query trends in different topics can be correlated, possibly with a certain time lag. We treat the query frequency data as a time series data stream. We conduct sample cross correlation analysis to identify whether there is a lag between trends that make them correlated.

5 Result and analysis

In this section we present the result and analysis.

5.1 Query pattern of academic searchers

We find very few queries (34 in the log) related to the medicine/physiology prize announcement. We suspect that academic searchers on this topic prefer other search engines such as PubMed. Therefore we neglect this topic in our analysis. For the other topics *chemistry* and *physics* we have collected 1489 and 1594 queries, respectively. The query trends are shown in Figure 2 and compared against the trend of all queries. For the two topics *chemistry* and *physics*, a growing trend is immediately observed following the prize announcements. It is also noted that both topics have a growing and oscillating trend initially, but a declining trend near the Christmas holidays, and then bounce back afterwards. Excluding the influence of the holidays, the query trend shows

an overall steady growth over time. For both topics, the initial growing trend of query frequency is in line with the global trend of all queries at the first glance, suggesting that academic queries increase at that specific time of the year. However, the trends are different after Christmas characterized by a bounce-back and a fall, and then oscillate, while the global trend is rather steady. This shows that queries of a specific topic may not be represented by the global trend and should be analyzed distinctively.
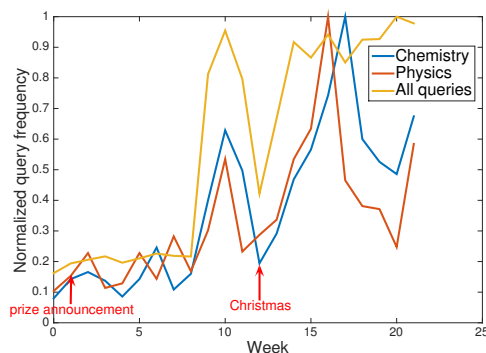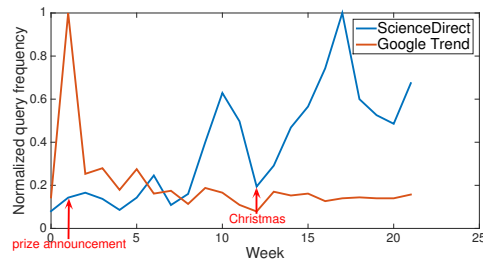

Fig. 2: Chemistry Prize.

Fig. 3: Query frequency of academic search starting from 1 week before the prize announcement. Frequency is normalized to fall between $[0, 1]$ for each topic and all queries respectively.
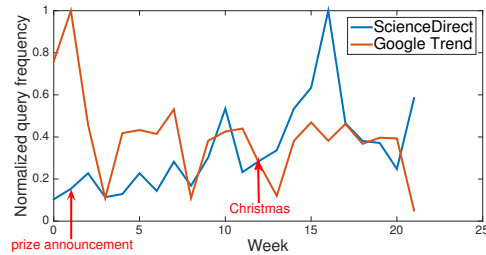
5.2   Academic search vs web search

In Figure 4 we compare the search trends between web search and academic search. The most prominent difference following the prize announcement is the surging spike in web search volume which is not seen for the academic search query volume. This shows that the major scientific events immediately arouse significant amounts of interests for users on web search engines, but not on academic search, suggesting that the event triggers "general interest" information needs rather than "in-depth" information needs for which an academic search engine might be a more appropriate source.

Besides the difference in the initial stage, contrary to the growing trend of academic search over time, the query trend of web search oscillates and declines. This shows that their user interests are only temporary and not consistent, for instance, news agencies and the general public that just want to capture and learn the news.

The distinct behavioral differences suggest that academic searchers, or at least academic search behavior, on web search engines are significantly outnumbered by other users, hence the different query trend. Correspondingly, we have only seen a weak correlation between the trends of these search engines (for *chemistry* and *physics* the Pearson correlation is -0.30 and -0.18, respectively). Due to the behavioral differences of

(a) Chemistry Prize.

(b) Physics Prize.

Fig. 4: Query frequency of academic searcher and web search. Frequency is normalized to fall between $[0, 1]$.

users on these search engines, it is not possible to predict the query trend of one search engine based on the other.

### 5.3 Query patterns of academic searchers and web searchers in different topics

In this section we consider behavioral commonalities for searchers with different topical interests. We examine the sample cross correlation [1] to identify lags between two query trends that might make them correlated.

Surprisingly, for two different topics, *chemistry* and *physics*, there is a moderate or high correlation between their query trends in academic search, as well as in web search. The results are shown in Figure 5. The sample cross correlation function indicates the correlation of trends with regards to the time lag (in days). Figure 5a shows that in academic search, the query trend of *chemistry* lags behind *physics* by about 8 days where the correlation is 0.56. Figure 5b shows a similar tendency in web search for the lag of chemistry, that the lag is only 1 day with the correlation being 0.63. Even when we exclude the impact of the Christmas holidays by leaving out the statistics from 2 weeks before the holiday, the correlation still holds at moderate or strong levels. This shows the behavioral commonality for users with different topical interests: their query trends in response to the occurrence of major scientific events are similar.
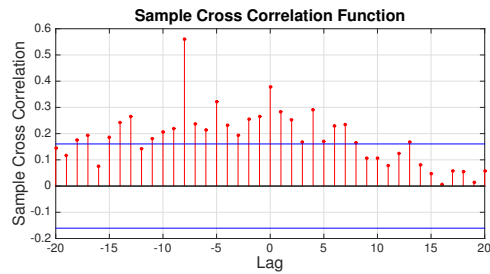
(a) Academic search.
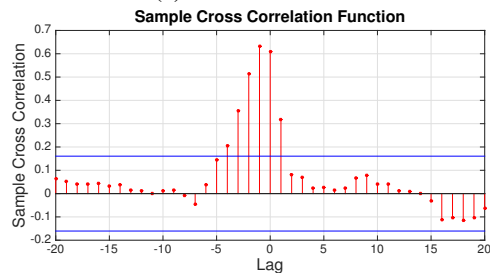
(b) Web search.

Fig. 5: Sample cross correlation function for *chemistry* and *physics*, in academic search and in web search respectively. Each trend in a search engine consists of daily query frequencies that are normalized between $[0, 1]$. This function indicates the correlation of different trends with regards to the time lag (in days). A negative lag indicates that *chemistry* lags behind *physics*. The confidence bound is at 0.1606 and -0.1606.

6 Implications

6.1 Theoretical implications

The theoretical implication from this work is that academic search engines receive an intrinsically different response from web search engines in response to major scientific events. This indicates distinct search behavior and information needs of academic searchers compared to users of a web search engine. Although academic search belongs to the broad category of web search, this specific domain should be examined carefully and the behavior insights gained in general web search, for instance concerning freshness of results, may not apply to academic search.

6.2 Practical implications

The practical implication is that when optimizing search engines to serve the information needs of users following the occurrence of major scientific events, different optimization measures should be taken for an academic search engine compared to a web search engine. For instance, since web search engine users have surging and temporary

interests, the freshness score of documents deserves a higher weight in time-sensitive document ranking models [2, 3, 8]. Academic searchers do not have this surging information need and their interests grow much slower, therefore such a factor needs to be weighed less importantly and different burst and decay models seem required.

7 Conclusion

In this paper we have examined the behavior of academic searchers in response to major scientific events through a log analysis. We have found that academic searchers exhibit a gradually growing search pattern, in contrary to the bursty and surging interests of the web searchers. Due to the difference in behavior between academic searchers and general web searchers, it is difficult to predict the search trend from one type of search based on the other. However, within academic search we see that searchers exhibit similar search patterns across topics, with a certain time lag. The same applies to web searchers. This demonstrates a certain pattern in the behavior of either groups.

This preliminary study reveals distinct behavior differences of academic searchers and web searchers and will benefit search engine design in the face of major scientific events. One limitation of this study is the small number of data points for observation, due to limitations of the transaction log available. Ideally the study should consider more "academic events" (e.g., other prize announcements, of importance to other fields of science and engineering), if sufficient data is accessible.

Another limitation of this study is the absence of analysis of individual searchers, e.g., whether their response to major events is in line with their long-term interests or just a temporary deviation. This is largely due to the limitation of the dataset, where in academic search we have not been able yet to collect enough user data, that is, of users who display a persistent search behavior during the whole period of observation. In our future work, based on more extensive user tracking, we aim to personalize trend analysis, examine trend prediction on a personalized basis, and offer personalized search results where appropriate.

Bibliography

[1] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. Wiley, 2015.

[2] S. Cheng, A. Arvanitis, and V. Hristidis. How fresh do you want your search results? In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1271–1280. ACM, 2013.

[3] N. Dai, M. Shokouhi, and B. D. Davison. Learning to rank for freshness and relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 95–104. ACM, 2011.

[4] H. Han, W. Jeong, and D. Wolfram. Log analysis of academic digital library: user query patterns. In *iConference 2014 Proceedings*, pages 1002–1008, 2014.

[5] B. M. Hemminger, D. Lu, K. Vaughan, and S. J. Adams. Information seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology*, 58(14):2205–2225, 2007.

[6] S. Jones, S. J. Cunningham, R. McNab, and S. Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3(2):152–169, 2000.

[7] H.-R. Ke, R. Kwakkelaar, Y.-M. Tai, and L.-C. Chen. Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect onsite in Taiwan. *Library & Information Science Research*, 24(3): 265–291, 2002.

[8] D. Lefortier, P. Serdyukov, and M. De Rijke. Online exploration for detecting shifts in fresh intent. In *Proceedings of the 23rd ACM International Conference on Conference on Information & Knowledge Management*, pages 589–598. ACM, 2014.

[9] X. Li and M. de Rijke. Do topic shift and query reformulation patterns correlate in academic search? In *The European Conference on Information Retrieval*, 2017.

[10] X. Li, R. Schijvenaars, and M. de Rijke. Investigating queries and search failures in academic search. *Information Processing & Management*, 53(3):666–683, May 2017.

[11] D. Lindberg. Internet access to the National Library of Medicine. *Effective Clinical Practice*, 3(5):256–260, 2000.

[12] X. Niu, B. M. Hemminger, C. Lown, S. Adams, C. Brown, A. Level, M. McLure, A. Powers, M. R. Tennant, and T. Cataldo. National study of information seeking behavior of academic researchers in the United States. *Journal of the American Society for Information Science and Technology*, 61(5):869–890, 2010.

[13] S. Pontis and A. Blandford. Understanding "influence:" an exploratory study of academics' processes of knowledge construction through iterative and interactive information seeking. *Journal of the American Society for Information Science and Technology*, 66(8):1576–1593, 2015.

[14] S. Pontis, A. Blandford, E. Greifeneder, H. Attalla, and D. Neal. Keeping up to date: An academic researcher's information journey. *Journal of the American Society for Information Science and Technology*, 68(1):22–35, 2017.