Towards Better Understanding of Academic Search

Madian Khabsa
Microsoft Research
Redmond, WA
madian.khabsa@microsoft.com

Zhaohui Wu and C. Lee Giles
The Pennsylvania State University
University Park, PA
zzw109@psu.edu, giles@ist.psu.edu

ABSTRACT

Academics have relied heavily on search engines to identify and locate research manuscripts that are related to their research areas. Many of the early information retrieval systems and technologies were developed while catering for librarians to help them sift through books and proceedings, followed by recent online academic search engines such as Google Scholar and Microsoft Academic Search. In spite of their popularity among academics and importance to academia, the usage, query behaviors, and retrieval models for academic search engines have not been well studied.

To this end, we study the distribution of queries that are received by an academic search engine. Furthermore, we delve deeper into academic search queries and classify them into navigational and informational queries. This work introduces a definition for navigational queries in academic search engines under which a query is considered navigational if the user is searching for a specific paper or document. We describe multiple facets of navigational academic queries, and introduce a machine learning approach with a set of features to identify such queries.

1. INTRODUCTION

Academic search engines have become the starting point for many researchers when they draft research manuscripts or work on proposals. Typically, there are two main retrieval systems that are used by academics. The first one is a citation database that is more of a traditional librarian search such as *Web of Science* and *Pubmed*, while the other is more similar to typical web search such as *Google Scholar*, and *Microsoft Academic Search*. Usage statistics tend to reflect user's preference for each type of systems. For example, 30% of Ph.D researchers relied on Google and Google Scholar as their main source for finding information in a survey conducted by the researchers of tomorrow project in 2012[1]. On the other hand, usage statistics from the University of California, Santa Cruz indicate that Google Scholar was used as a secondary source of information rather than a primary one in 2010 [4].

The need for retrieving relevant information in scientific domains has lead to many contributions in information retrieval [12]. For example, the idea behind indexing documents using keywords have originated from the field of librarianship [12]. Similarly, the intuition behind *PageRank* goes back to citation indexing, and using the number of citation as a proxy for measuring importance.

In this work we focus on academic search systems that use keyword base search. We study user behavior of an academic search engine using query logs of three years. By observing user query patterns, we were then motivated to study the user query intent by classifying academic search queries into navigational and informational queries. Search engine queries have typically been classified into three main categories [2]: 1) Informational: the user is seeking some information available on the web; 2) Navigational: the user is trying to reach a particular website; and 3) Transactional: where the user is trying to perform some type of transaction. Query intent classification is an important part of any search engine because the query type affects how the query is handled. Identifying navigational queries is essential for devising specific ranking functions that rank navigational queries only. Similarly, if a query is known to be navigational, the search engine may choose to present the results page differently by either showing a single result or few results. In this work we introduce the concept of academic navigational query, and define multiple facets through which a query should be considered navigational. To the best of our knowledge, this is the first work that studies academic query classification.

2. RELATED WORK

Although there is a rich literature related to academic search in both information retrieval and data mining community, few work studies the usage of academic search engines and seeks to understand academic search using real world user query data. We briefly review the related research on academic search in the following directions.

Existing academic search engines. Earlier academic search engines that build upon automatic citation indexing and metadata extraction include Citeseer [3], followed by Arnetminer which focuses more on extraction, analysis and mining of academic social network and providing academic rankings [14]. In the industry, two typical instances of

[1]http://www.webarchive.org.uk/wayback/archive/20140614040703/http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow.aspx

JCDL '16, June 19–23, 2016, Newark, NJ, USA.

DOI: http://dx.doi.org/10.1145/2910896.2910922

Table 1: Search type percentage. Document and author search

Search Type	Percentage
Document	92.73%
Author	6.9%
Other	0.2%
Total	100%

academic search engines are Google Scholar and Microsoft Academic Search.

Query type identification. Most query type identification has focused on general Web search engines, with different methods applied in different settings. For example, Kang and Kim used the difference of distribution, mutual information, the usage rate as anchor texts, and the POS information for the classification [7]; Jansen et. al. proposed a rule based method based on a set of heuristics [5]; and Lee et. al. studied user-click behavior and anchor-link distribution [9]. The closest work to ours is Kan and Poo [6] where they focused on detecting known item search in online public access catalogs. However, our work introduces the concept of navigational queries and formally categorizes the cases that belong the concept.

Personalized academic search. Some research also investigated the personalized academic search by user modeling based on query log [13].It is worth noting that traditional user models such as collaborative filtering that do not consider document content features may fail since they play a more important role in ranking document than user similarity.

3. ACADEMIC QUERY LOGS

Our work was motivated by examining the usage behavior of academic users of CiteSeerX, which indexes publications in computer science and engineering, physics, and economics. The search engine provides multiple search types, most notably document search and author search. The search sessions received by the search engine between September 2009 and March 2013 are used in this study.

The proportions of search types are presented in Table 1. As seen in the table, the majority (92.73%) of search sessions belong to document search. In addition, there is a reasonable interest in searching for author names with 6.9% of the search requests targeting the authors index. We conjecture two possible reasons for this. In the first scenario users are interested in searching for papers by author name assuming they remember one of the authors of a paper but do not recall the title. The second reason is that individual authors are searching for themselves to track down the number of citations they received, the number of papers they have being indexed by the search engine. Similarly, tenure and promotion committees would use the author search to track down the body of work for a given scholar. An important feature of queries is query length, which might to some extent indicate the user query intent. For example, users issuing longer queries might be more likely searching for more specific information. By examining the queries with type *document search* we found that the average length of a query is 4.76 terms.

4. QUERY TYPE CLASSIFICATION

Most of query intent identification has focused on general search engines with many approaches being applied [7, 5, 9]. However, there has not been any work that identifies navigational queries in academic search, to the best of our knowledge. In fact, we are not aware of any work that categorizes query intent in the academic search. Perhaps because traditional library search was the defacto standard in academic search until recently. The ease of use of current academic search engines such as Google Scholar, combined with their similarity to traditional web search that most people have become familiar with creates an opportunity to study the query intent of the users.

In academic search it is possible to categorize queries into at least two types: *navigational* and *informational*. It is not straightforward how to define transactional queries, if they exist, in the academic setting. Providing a complete taxonomy of query intent in academic search is beyond the scope of this work. Rather, we focus on identifying navigational queries. We define a **navigational query** as a query for which the user is looking for a specific scholarly document, which can be a paper, book, thesis, etc. Correctly identifying navigational queries is important, because rankers are heavily influenced by citations which can lead to highly cited papers being ranked higher than the target paper if the target paper is new or not well cited. Furthermore, papers whose title contains general terms are more likely to be susceptible because there are large number of matches. There are multiple facets for navigational queries in academic search. For example, a user might look for a given document by:

- Document Object Identifier (DOI): *10.1038/nature14106*
- Full title query: *The Google file system*
- A combination of an author and title information: *jeff dean mapreduce*
- Author and year/venue information: *leskovec 2009 news cycle*
- Author names for a well known work: *Cormen Leiserson Rivest Stein*, or *hopcroft motwani ullman*
- A combination of author names along with some paper's distinguishing terms: *dic brin motwani*

In the first scenario when users search by DOI, it is sufficient to check if the query matches a database of DOIs or not. However, other cases are not as trivial. For example, title queries are not easily detectable. First of all, extracting titles from papers is not always accurate. In addition, although a query can be checked against a list of titles, there are many short and ambiguous queries that might match at multiple title positions. On top of that, no search engines contain all academic documents [8], hence identifying a title navigational query that the search engine does not have a result for it may be used as signal to locate this missing document. However, other cases can be more subtle to identify. For example the following queries that were found in the logs of the academic search engine are not as obvious. In the query *leskovec 2009 news cycle* there exists an author name and a year along with subset of a title that would identify

Table 2: Navigational Query Features

Feature	Description
#_tokens	number of tokens in a query
has_year	whether a term in the query matches a regex for identifying year
has_stop_word	whether the query has stop words
has_punctuation	whether the query has punctuation
#_authors	the number of tokens in the query identified as author name
author_ratio	the ratio of query terms identified as author names to the query length
is_title_match	whether the query matches a title in the search engine's index

the work.[2] Similarly, *dic brin motwani* and *lift brin motwani ullman* both refer to the dynamic itemset counting paper by S. Brin and R. Motwani [3]. Finding the correct matching of those queries might need more sophisticated approach than simple rules.

4.1 Approach

The problem is modeled as a binary classification problem. Given a query q, we would like to classify it into one of the following classes {`Navigational, Informational`}. Each query q is represented as a vector of the features described in Table 2. The features are crafted to capture the multiple facets that represent navigational queries. For example, `#_tokens` is chosen after noticing that many navigational queries have more terms than informational queries because they contain a title. On the other hand, `is_title_match` can be a good signal in general, but if the query term is general, a match does not necessarily make the query navigational. Other syntactic features such as `has_stop_word` , and `has_punctuation` are aimed at identifying title queries. The intuition is that users rarely use such terms in informational queries, and they are more likely to be part of title.

As shown in the examples, the mention of author names is one facet of navigational queries. However, it is not always the case that a mention of an author means a navigational query. For example, the following query that was found in the logs of the academic search engine is not considered navigational: `mccallum nigam` because these two authors have coauthored more than one paper together, and this query cannot be interpreted to refer to a single paper. Nevertheless, the presence of an author name is one of the indicators of navigational queries. Therefore, we create a feature to represent the number of query tokens that are identified to be an author name. Identifying whether a token refers to an author name or not was not as trivial as checking against a dictionary of all possible names. Initially we started by using the author list of DBLP[4] as a names dictionary assuming that it would have low false positive rate since it is manually curated. However, the false positive identification was still high as tokens such as `network` matched as an author.

Therefore, we adapted a language model approach to identify author names. For every token t we estimate three probabilities: $P(t|author)$, $P(t|title)$, and $P(t|abstract)$ where author, title, and abstract refer to the token appearing in the author, title, or abstract section of the paper, respectively. We estimate each of the probabilities for every token over all the fields in the academic search engine's index. A token t is considered an author iff:

$$P(t|author) > P(t|abstract) \wedge P(t|author) > P(t|title)$$

Gradient Boosted Trees (GBT) are used to train a classifier for identifying navigational queries. The number of stumps and the learning rate parameter are chosen using grid search over the range $[10, 400]$ and $[10^{-4}, 10^{-1}]$, respectively. SMOTE oversampling is used to oversample the navigational queries because the dataset is imbalanced.

4.2 Dataset

To build the dataset, we first randomly sample 1000 queries from the user search logs and then keep only the queries in document search type, which results in 553 in total. However, notice that this small number of samples might not give reasonable coverage of all possible positive samples (navigational queries), we did multiple rounds of sampling and use the aforementioned heuristics to match possible positive candidates. The dataset was then augmented by those positive examples that might not have enough presence in the randomly sampled dataset, such as examples with author names. We also added comparable number of negative examples to counter for that effect. At the end, the dataset contained 579 queries. Each query was manually inspected by two human judges and tagged as either navigational or informational. When the judges had mismatching labels, they conferred and agreed on a mutual tag. In the manually tagged queries, 12.5% were found to be navigational.

4.3 Experiments

The performance of the classifier is shown in Table 3. The numbers are for a 5 fold cross validation, with the training fold being randomly split into 90:10 with the 10% used to validate the grid search parameters. Oversampling using SMOTE was only conducted on the training fold, with the test fold remaining untouched. We compared the performance of the boosted tree classifier with that of an SVM with RBF kernel, and with that of a random forest. All parameters for both baseline classifiers are configured with grid search, similar to the GBT. The highest precision, and overall F score was attained with GBT as can be seen in table 3. The numbers in the table refer to the average precision, recall and F score obtained through the 5 fold cross validation, with the standard deviation reported between parenthesis. The importance of each of the features is shown in Figure 1. The number of tokens within a query is the most important feature, which can be explained by title queries that tend to have higher number of tokens. Similarly, the title match feature which is closely related to the number of tokens in

[2]Meme-tracking and the dynamics of the news cycle (KDD'09).

[3]Dynamic itemset counting and implication rules for market basket data (SIGMOD'97).

[4]http://www.dblp.org/search/index.php

Table 3: Navigational query classification performance for multiple learning algorithms. Numbers between parenthesis refer to standard deviation in 5 fold cross validation

Method	Precision	Recall	F1
GBT	0.68 (0.03)	0.68 (0.09)	0.677 (0.04)
SVM (RBF)	0.67 (0.05)	0.63 (0.12)	0.64 (0.07)
Random Forest	0.71 (0.06)	0.59 (0.14)	0.62 (0.09)

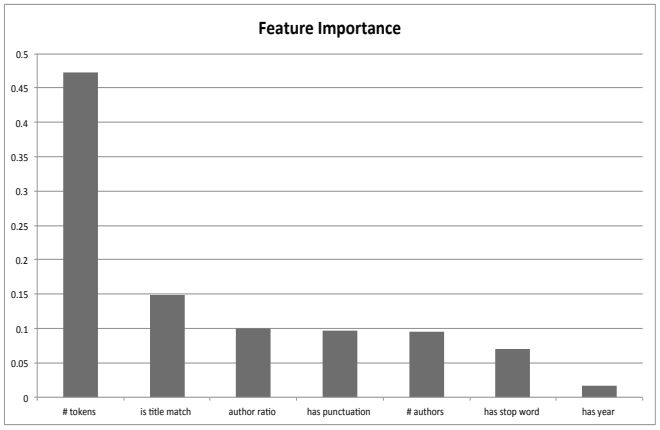

Figure 1: Feature importance of navigational query classification.

the query ranks second in terms of importance, followed by author ratio feature.

It is worth noting that classifying navigational queries is notoriously a hard task in the web domain, and it would be at least as hard within the academic realm. For example, Jansen et al.[5] were only to obtain 74% for overall web query intent classification, which is not limited to navigational. Others were able to obtain 70% precision with high recall for informational queries [1]. Many studies for navigational query classification at web search engines have relied on click through rates [9, 11, 10]. While these methods were effective overall, they remain passive and depend on the presence of queries and clicks in the logs to be able to accurately classify them. This presents a challenge when new queries that have not been seen before arrive, or when users refer to an academic paper using a new combination of keywords.

5. CONCLUSION AND FUTURE WORK

We studied academic search based on user query logs. We then introduced the concept of academic navigational query and studied the problem of academic query type classification on a new dataset with human judgments. We proposed a set of features to learn the classifier. The results showed the effectiveness of the proposed features and demonstrated the challenge of the problem.

One of our ongoing work is on implementing our new academic ranking methods that take advantage of both navigational query classification and the learned academic document rank functions. We could add the navigational query classification result as a new feature for the ranker or train separate rankers for different types of queries.

6. ACKNOWLEDGMENTS

We acknowledge partial funding by the National Science Foundation.

7. REFERENCES

[1] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind web queries. In *String processing and information retrieval*, pages 98–109. Springer, 2006.

[2] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[3] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98. ACM, 1998.

[4] C. Hightower and C. Caldwell. Shifting sands: science researchers on google scholar, web of science, and pubmed, with implications for library collections budgets. *Issues in Science and Technology Librarianship*, (63):4, 2010.

[5] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, pages 1149–1150. ACM, 2007.

[6] M.-Y. Kan and D. C. Poo. Detecting and supporting known item queries in online public access catalogs. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, pages 91–99. IEEE, 2005.

[7] I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR*, 2003.

[8] M. Khabsa and C. L. Giles. The number of scholarly documents on the public web. *PLOS one*, 9(5):e93949, 2014.

[9] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 391–400. ACM, 2005.

[10] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2008.

[11] Y. Lu, F. Peng, X. Li, and N. Ahmed. Coupling feature selection and machine learning methods for navigational query identification. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 682–689, 2006.

[12] M. Sanderson and W. B. Croft. The history of information retrieval research. *Proceedings of the IEEE*, 100(Special Centennial Issue):1444–1451, 2012.

[13] Y. Sun, H. Li, I. G. Councill, J. Huang, W.-C. Lee, and C. L. Giles. Personalized ranking for digital libraries based on log analysis. In *Proceedings of the 10th ACM Workshop on Web Information and Data Management*, WIDM '08, pages 133–140, 2008.

[14] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.