The survey of large-scale query classification

Cite as: AIP Conference Proceedings **1834**, 040045 (2017); https://doi.org/10.1063/1.4981641
Published Online: 28 April 2017

Sanduo Zhou, Kefei Cheng and Lijun Men

The Survey of Large-scale Query Classification

Sanduo Zhou [1, a)], Kefei Cheng [1, b)] and Lijun Men [2)]

[1]*College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China;*
[2]*Chongqing University of Posts and Telecommunications, Chongqing, 400065, China.*

[a)] Corresponding author: 726679904@qq.com
[b)] chengkf@cqupt.edu.cn

Abstract. In recent years, a lot of researches have been done on query classification. The paper introduces the recent researches on query classification in detail, mainly including the source of query log, the category systems, the feature extraction methods, classification methods and the evaluation methodology. Then it discusses the issues of large-scale query classification and the solved methods combined with big data analysis systems. The research result shows there still are several problems and challenges, such as lack of authoritative classification system and evaluation methodology, efficiency of the feature extraction method, uncertainty of the performance on large-scale query log and the further query classification on the big data platform, etc.

Key words: Query Classification, Feature Extraction, The Methods of Query Classification, Query log.

INTRODUCTION

With the development of web technology, the internet has stepped deeply into every aspect of people's life and became the main source to acquire the information. In the era of information explosion, a growing number of users acquire and publish information on the internet. Search engine has become the main tool for people to obtain the desired information. It collects information from internet according to certain strategies and specific computer programs and provides users with retrieval service after organizing and processing the information. However, most of the search engines adopt the search method based on the combined keywords, and the number of input words is limited. Therefore, most users cannot use few keywords to accurately summarize their demand for information. This situation leads the search engine to return the irrelevant information in most times. In order to return the correct information to users and improve the quality of the search engine, it is urgent for search engines to understand the users' query intents. Query classification is an effective way to understand user intents, which can further improve the quality of web search. In recent years, both the domestic and the foreign researchers have done a lot of effort on studying how to infer the intents or interest of search queries via search engine query log. Query log of search engine records the users' query behaviors and reflects the user query intents and requirements, which can be used to improve the performance of search engine. Therefore, query log becomes the data source of the researches on query classification.

Recently, the search engine develops rapidly and the amount of users is increasing year by year. So the query log is increasing at rapid speed not only in size but also in variety. Log data generated from Facebook exceed more than 300 TB and the amount of data that Baidu[1] Company dealing with reach to the dozens of PB every day. "Big data" has become a challenge in the field of query classification. Web search engines (Google, Amazon, and Yahoo) are the first to face the problem of big volume of data to handle in real time [1]. Management tools have been developed

2017 5th International Conference on Computer-Aided Design, Manufacturing, Modeling and Simulation (CDMMS 2017)
AIP Conf. Proc. 1834, 040045-1–040045-7; doi: 10.1063/1.4981641
Published by AIP Publishing. 978-0-7354-1504-1/$30.00

to solve the problems of big data, such as Hadoop, Spark, etc. These tools can be used for large-scale query classification. However, there are still few researches on large-scale classification.

The rest of this paper is organized as follows. The existing category systems of query classification are described in Section 2. Section 3 has presented the query classification methods and the evaluation methods. Large-scale query classification is discussed in Section 4. Finally, conclusion will be put in the Section 5.

CATEGORY SYSTEMS

The goal of query classification is to classify the users' queries into the predefined categories according to queries' topics or user intent. A lot of researches on the category system of query classification are done to map search queries into categories, which are mainly based on intent and topic query. Early in 1997, Schneiderman considered that the internal information demand makes people began to use the information retrieval system, and the intent of queries is to get the information. Therefore, the authors considered all the queries as informational queries. However, Broder, et al. [2] found that the intent of queries is not only the purpose of reading information on static web page after analyzing the query log of AltaVista. Users like to visit the websites which are known by them and the websites which they want to have further interaction. So they classified the intent of queries into "Informational", "Navigational" and "Transactional". In [3], the authors thought that the "Transactional" cannot sum up all the resource in the internet and suggested changing "Transactional" into "Resource". "Resource" is no longer limited to the general web activity, but any available resources of websites (not informational). They proposed more detailed hierarchy structure based on the "Resource". Their category systems are widely adopted and many researches of category are studied based on Border's and Rose's category system in the future. Table 1 shows the categories of different category systems based on query intent in recent years.

TABLE 1. Category system of query intent classification.

Author	Classification system
Broder [2]	Navigational/Informational/Transactional
Rose [3]	Navigational/Informational/Resource
Kang & Kim [4]	Topic Finding/Homepage Finding/Service Finding
Marchionini [5]	Lookup/Learn/Investigate
Mendoza [6]	Informational/Non-informational/Ambiguous
Waller [7]	Navigational/Informational/Transactional/Leisure

Query topic classification tries to map the search queries into a predefined topic category, such as Leisure, News and Business, etc. Spink, et al. [8] divided the queries into 11 categories, such as "Leisure", "Gender", "Economic", "Internet", etc. In [9], the authors set up 10 basic categories and 2 subsidiary categories of Japan-referring query log. Queries are divided into 10 topical categories according to the query topic classification of Sogou[2] query logs in Liu's research [10]. Many researchers have put forward their own classification system, and the related experiments are done to verify their validity. These category systems differ in thousands ways, but a standard category system has not been formed yet.

QUERY CLASSIFICATION METHODS

Features Extraction Methods

Now, the query log dataset which researchers use to study query classification mainly comes from the current open search engine, such as AltaVista, Excite, AOL and Sogou, which are shown in table 2.

TABLE 2. Several open query log.

Source	Sogou	AOL	AltaVista	Excite
Language	Chinese	English	English	English
Date	2008.06	2006.03-05	2002.12	2001.04
Queries Number	51,537,393	36,389,567	7,175,648	-
Average Length of Query	3.3	2.35	2.35	2.4

The average length of English query is 2.35, and the average length of Chinese query is 3.3. In a query log, it is obvious that the query string contains few words, and the features are sparse. How to obtain sufficient features from simple and sparse query log is an important topic of query classification. Query feature extraction methods are divided into three types based on [11]. The characteristics of these methods are compared in table 3.

TABLE 3. Feature extraction method.

Methods	Dimensions	Characteristic
Feature Extraction Based on the Query Expression	Query Words	Sparse feature, Low cost, Low accuracy, Rich timeliness
	Query Frequency	
	Query Length	
Feature Extraction Based on the Retrieved Results	Anchor-link	Rich features, High cost, Complex operation, Poor timeliness
	Web Advertising	
	Wikipedia	
	Pseudo Relevance Feedback	
	URL Topic	
	URL-key	
Feature Extraction Based on Users' Behavior	Users' Interactive Behavior	High accuracy, Rich feature, Complex Operation, Poor timeliness
	Context Change	
	Users' Click Behavior	

In order to obtain higher accuracy of query classification, some researchers have studied the query classification from the multidimensional web query data. Jiang, et al. [12] extracted classification features from user-submitted queries to form a feature vectors, such as the length of query term, the ranking of the query results in the same session and the number of queries in the same session, etc. Then they used SVM to classify the intent of users' query into information or non-information. In [13], the authors proposed a classification model of query log based on topics and classified the queries from multi-dimensional features, such as the URL dimension, Session dimension and query terms dimension of the query expressions. These methods further enrich the query features and improve the accuracy of classification, but they also increase the complexity of the operation.

Query Classification Methods

At present, query classification is divided into artificial classification and automatic classification according to whether the human beings participate in classification or not. Their advantages and disadvantages are listed in table 4.

TABLE 4. Advantages and disadvantages of different classification methods.

Query classification method			Characteristic			
			Precision	**Recall**	**Artificial Cost**	**Time Cost**
Artificial Classification			Higher	Low	Higher	Higher
Automatic Classification	Rule-based Classification		Low	High	High	Lower
	Machine Learning Classification	Supervised Learning	High	High	High	Lower
		Unsupervised Learning	Low	High	Lower	Lower
		Semi-supervised Learning	High	High	Low	Low

The precision of artificial classification is higher than others. However, it is difficult for artificial classification to deal with the low frequent queries because of its long tail effect, resulting in low recall. Law, et al. [14] developed a game and completed the query classification when users play this game. This method has reduced the cost of manual classification in a certain extent. Researchers try to use the automatic query classification to improve the recall and reduce the artificial cost and time cost. Machine learning classification is a commonly used method in automatic classification, which is classified into supervised learning, unsupervised learning and semi-supervised learning.

Supervised learning uses the given category of samples to adjust the parameters of classifier and obtains an optional model which is trained to reach to all the required performance. Then all the input is accordingly mapping for the output by using this model. Supervised learning can achieve the purpose of query classification after making simple judgments of the output. Kang, et al. proposed a user query scheme. It takes the labeled data which is a part of TREC-2000 as training sets and uses the distributional difference of distribution of query terms, mutual information, the usage rate as anchor texts, and the POS information for the classification [4]. Liu, et al. [15] used the typical decision tree algorithm combining nCS, nRS and distribution of click for classification. Their experimental classification results are better than Lee's.

Unsupervised learning is usually applied in query classification without the given category. Sun, et al. [16] proposed a hierarchical clustering algorithm for large-scale query log. The algorithm can accomplish clustering by using different similarity calculation methods and obtain the cohesive categories of queries. Ren, et al. [17] improved the Sun's algorithm and use for clustering on the query log. The artificial cost and time cost of unsupervised learning is lower without labeled data. It classifies the queries based on similar degree, but similarity may not reflect the relationship among the specific category systems. It is hard for classification results to map to the predefined category systems.

In practical issues, there is usually only a small amount of labeled data and the cost of tagging data is very large. Therefore, semi-supervised learning can generate the appropriate classification function by using a small number of labeled samples and a large number of unlabeled samples. Zhang, et al.[18] structured a model for query classification based on semi-supervised learning. It uses the external network knowledge library to extend the features of short text. The initial classifier is used to achieve the full use of the unlabeled samples by iterative self-learning, which can solve the bottleneck problem of tagging and improve the performance of the classifier.

The Evaluation Methods of Query Classification

There is no specific evaluation method in the field of query intent classification, and the evaluation methods of query classification results usually reference the evaluation methods of text classification. Relevant indicators include accuracy, recall, precision, error rate, and the value of F_1 are used to evaluate the performance of a single category. At present, macro averaging and micro averaging are used to measure the performance of classification methods in all categories. Macro averaging is the arithmetic average of performance indicators for each class, while micro averaging is the arithmetic average of the performance metrics for each query. The indicator of micro averaging evaluation is susceptible to the classification performance of the largest categories while the indicator of macro averaging evaluation is susceptible to the classification performance of the smallest categories.

LARGE-SCALE QUERY CLASSIFICATION

With the development of internet technology, a growing number of web users use a variety of search engines to obtain the information. And the amount of internet information has grown rapidly not only in size but also in variety. This change has great effect on users' queries intents. It is harder for search engines to classify the large-scale queries into different categories. Beitzel, et al. [19] proposed a method for classification without using external sources of information, which combining manual classification, supervised learning classification, and rule classification. Comparing with the single method, the combined method can obtains superior results of classification. In order to address the large-scale classification, Sun, et al. proposed an algorithm named Chimera. This algorithm use a combination of machine learning, hand-crafted rules, developers, analysts, and crowd workers to achieve accurate, continuously improving, and cost-effective classification [20]. Chimera can successfully classify ten millions of products into more than 5000 categories at Walmart Labs. In [21], a new coding scheme called Tree Quantization is proposed to provide a combination of high accuracy and fast encoding which is attractive for large-scale retrieval and classification systems.

The above methods can solve the problem of large-scale query classification in a certain extent. However, the scale of data which they can address with a single computer is limited by the storage and computational performance. And the volume of data is enormous right now and predicted to reach 35 ZB by 2020[22]. Web log is growing rapidly as well. This large data was termed as big data. Big data is characterized by high dimensionality and large size. These features mainly rise following issues which also exist in large-scale query classification:

Large-scale data may create issues such as heavy computational cost and algorithmic instability [23].

The larger the dataset to be processed, the longer it will take to analyze [24].

The traditional machine learning classification cannot process the large-scale query log more quickly than before. Therefore, big data analysis, such as Hadoop [25-26] and Spark [27], systems have flourished to solve the problems. Few researchers have done analysis researches of the query log based on the big data platform. Wei, et al. [25] established a new processing system for remote parallelization analysis based on Hadoop. This system reduced the performance bottleneck of computing power and storage capacity, thereby saving much time and improving efficiency significantly. In [26], a novel scalable distributed system is proposed, which combines a MapReduce based platform with NoDB paradigm. Their experimental analysis shows that it can significantly reduce the data-to-query latency with respect to comparable state-of-the-art distributed query engines, like Shark, Hive and HadoopDB. Liu, et al. presented a simple and complete system for sentiment mining on large-scale datasets using a Naïve Bayes classifier with the Hadoop framework which can scale up easily without a database [28]. The above mentions can be used in large-query classification to speed up the computation and improve algorithmic instability of query classification in the future.

CONCLUSION

In this paper, we have illustrated the category systems of query classification and the feature extraction methods. Then the classification methods and the evaluation methods are introduced in detail. Researches on query classification have made great progress after years of effort, but the following questions and challenges still remain in query classification: lack of authoritative classification system and evaluation methodology, efficiency of the feature extraction method and uncertainty of the performance on large-scale query log. In the past, the amount of user data for classification researches is small-scale. But with the development of internet, the scale of user query log has become more and more large. Query classification is facing the challenge of big data. We have discussed the issues which are existed in large-scale query classification and the researches to solve these issues by combining machine learning with big data analysis in this paper. Although few researchers have done their effort on large-scale classification, the performance of different classification methods on large-scale query log still cannot be evaluated. How to classify the large-scale queries combining the classical query classification method with the big data processing technology, especially in Hadoop or Spark, will be a major trend in the future research of query classification.

ACKNOWLEDGMENTS

This work is partially supported by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant Nos. KJ1400441 and KJ120511. It is also supported by Demonstrative Project

Supported by Science and Technology of Chongqing Science and Technology Commission under Grant Nos. CSTC2014zktjccxyyBX0023. The authors are thankful to the referees for their valuable comments and helpful suggestions.

REFERENCES

1. Jamiy F E, Daif A, Azouazi M, et al. The potential and challenges of Big data - Recommendation systems next level application. J. International Journal of Computer Science Issues, 2014.
2. Broder A. A taxonomy of web search. C. //ACM Sigir forum. ACM, 2002, 36(2): 3-10.
3. Rose D E, Levinson D. Understanding user goals in web search. C. //Proceedings of the 13th international conference on World Wide Web. ACM, 2004: 13-19.
4. Kang I H, Kim G C. Query type classification for web document retrieval. C. //Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2003: 64-71.
5. Marchionini G. Exploratory search: from finding to understanding. J. Communications of the ACM, 2006, 49(4): 41-46.
6. Mendoza M, Baeza-Yates R. A web search analysis considering the intention behind queries. C. //Web Conference, 2008. LA-WEB'08, Latin American. IEEE, 2008: 66-74.
7. Waller V. Not just information: Who searches for what on the search engine Google? J. Journal of the American Society for Information Science and Technology, 2011, 62(4): 761-775.
8. Spink A, Wolfram D, Jansen M B J, et al. Searching the web: The public and their queries. J. Journal of the American Society for Information Science and Technology, 2001, 52(3): 226-234.
9. Buzikashvili N. Query Topic Classification and Sociology of Web Query Logs. J. Computación Sistemas, 2015, 19(4): 633-646.
10. Liu F, Li Y, Lv X Q, Li Z. Research on Query Topic Classification Method. J. New Technology of Library and Information Service, 2015, 31(4):10-17.
11. Lu W, Zhou H A, Zhang X. Review of Research on Query Intent. J. Journal of Library Science in China, 2013, 39(1): 100-111.
12. Jiang Z L, Zhang H Q, Automatic Identification of Query Intent by Using Multiple Features. J. Computer and Information Technology, 2015, 23(1): 1-4.
13. Jiang D, Leung K W T, Ng W. Query intent mining with multiple dimensions of web search data. J. World Wide Web, 2016, 19(3): 475-497.
14. Law E, Mityagin A, Chickering M. Intentions: A game for classifying search query intent. C. //CHI'09 Extended Abstracts on Human Factors in Computing Systems. ACM, 2009: 3805-3810.
15. Y, Zhang M, Ru L, et al. Automatic query type identification based on click through information. C. //Asia Information Retrieval Symposium. Springer Berlin Heidelberg, 2006: 593-600.
16. Sun R, Jin P. Hierarchical Clustering Method for Large-scale Chinese Query Logs. J. Bulletin of Science & Technology, 2012.
17. Ren Y W, Lv X Q, et al. Hot Query Content Extraction in Search Engine Logs. J. Computer Applications and Software, 2015, 32(12):16-21.
18. Zhang Q, Liu H L. An Algorithm of Short Text Classification Based on Semi-supervised Learning. J. New Technology of Library and Information Service, 2013(2):30-35.
19. Beitzel S M, Jensen E C, Lewis D D, et al. Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs. J. Acm Transactions on Information Systems, 2007, 25(2):107-108.
20. Sun C, Rampalli N, Yang F, et al. Chimera: Large-scale classification using machine learning, rules, and crowdsourcing. J. Proceedings of the VLDB Endowment, 2014, 7(13): 1529-1540.
21. Babenko A, Lempitsky V. Tree quantization for large-scale similarity search and classification. C. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4240-4248.
22. 22. Goli-Malekabadi Z, Sargolzaei-Javan M, Akbari M K. An effective model for store and retrieve big health data in cloud computing. J. Computer Methods and Programs in Biomedicine, 2016, 132: 75-82.
23. Fan J, Han F, Liu H. Challenges of big data analysis. J. National science review, 2014, 1(2): 293-314.
24. Bhardwaj V, Johari R. Big data analysis: Issues and challenges. C. // International Conference on Electrical, Electronics, Signals, Communication and Optimization. 2015.

25. Wei Y Q, Zhou G G, Xu D, et al. Design of the web log analysis system based on Hadoop. C. //Advanced Materials Research. Trans Tech Publications, 2014, 926: 2474-2477.
26. Tian Y, Alagiannis I, Liarou E, et al. DiNoDB: Efficient large-scale raw data analytics. C. //Proceedings of the First International Workshop on Bringing the Value of Big Data to Users (Data4U 2014). ACM, 2014: 1.
27. Gong J, Analyzing Users' Wab log Based on Spark. J. Guangdong Communication Technology 2015, 35(1):16-21.
28. Liu B, Blasch E, Chen Y, et al. Scalable sentiment classification for big data analysis using Naive Bayes Classifier. C. //Big Data, 2013 IEEE International Conference on. IEEE, 2013: 99-10