UNDERSTANING THE CONTEXT
IN INFORMATION RETRIEVAL

by

Peilin Yang

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Engineering

Fall 2016

UNDERSTANING THE CONTEXT

IN INFORMATION RETRIEVAL

by

Peilin Yang

Approved: _____

Xxxx Xxxx, Highest Degree
Chair of the Department of Xxxx

Approved: _____

Xxxx Xxxx, Highest Degree
Dean of the College of Xxxx

Approved: _____

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Xxxx Xxxx, Highest Degree
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Xxxx Xxxx, Highest Degree
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Xxxx Xxxx, Highest Degree
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Xxxx Xxxx, Highest Degree
Member of dissertation committee

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Information Retrieval (IR) is one of the most evolving research fields and has drawn extensive attention in recent years. In IR research, context plays a crucial role and it is the fundamental basis of some very important IR research efforts. In this proposal, we identify the contexts of three different domains of IR researches and investigate their impacts.

The first scenario is the unified IR evaluation system. Previously, the typical work-flow of proposing a new IR ranking model is to implement the algorithm using the most convenient tool (e.g. Indri) of the researcher's choice. The ranking model is then evaluated against standard data collections and the performances are compared with commonly used baselines (e.g. BM25) and the advantage of the proposed model is shown. The problem of such methodology is that different tools and experimental settings may result in different results for the same model and thus cast doubts on the real effectiveness of the proposed ranking models. We have monitored the adverse effect brought by the aforementioned problem and explore the usage of the standardized IR evaluation environment. VIRLab as our first attempt to alleviate the problem. It provides an easy-to-use IR evaluation system where user of the system can implement her/his our model without considering too much about the underlying framework. A web based Reproducible IR Evaluation System (RISE) is the first ever known system that users of the system can collaborate with each other and thus make the normalization of baseline models much more easier and thus can be trusted by future researchers. A large scale experiment based on RISE serves as the the performance reference of most widely used IR models.

Another example is the boundary theory of ranking model performance. The context of this line of research is the bag-of-terms representation of the document

assumption and the widely used statistics, e.g. Term Frequency (TF) and Inverted Document Frequency (IDF) that the ranking model consists of. We first compare the optimal performances of state-of-the-art ranking models and find the optimums are similar for those models even the underlying theories are different. To dig it deeper we use the cost/gain analysis which is commonly used in learning-to-rank (LTR) regime to find the optimum of single term queries for a family of ranking models that share similar strategy of combining key signals. The result shows that although the performances of state-of-the-art ranking models are quit close to the practical optimum there is still some room for improvement.

The third domain is to investigate the usage of opinion to contextual suggestion. It is commonly agreed that how to model user profile is the key to the solution. We first model the user profile using the venue's category and description of user activity history. We further improve the method by leveraging the opinions from user's activity history to model the user profile. Such methodology naturally utilizes the rich text associated with the users which naturally makes the problem as a retrieval problem. By modeling the candidate suggestions in the similar fashion, the similarities between the candidates and the user profile are used to score the candidates. Experiments on TREC collections and a Yelp data collection reveals the advantage of our method.

For the future work, there are mainly two directions we would like to explore more deeper. The first one is to quantify the impact of the context of unified IR evaluation system. There is no previous work on how much difference does the usage of different retrieval tools bring. We hope to be the first to report on standardizing and quantifying the impact so that the IR community could be aware of such divergence and can better evaluate the contributions of using various tools. The second one is to provide more sound justification about the boundary theory of ranking model performance. Specifically, we want to extend the current analysis on the single term queries to multiple term queries. We also would like to try other method, e.g. explanatory analysis, to achieve the same goal.

Chapter 1

INTRODUCTION

The past decades have witnessed the tremendous success of World Wide Web. People all over the world can now access to publicly available information via commercial search engines such as Google, Microsoft Bing with great ease. According to the online statistics [1], Google now (as of October 2016) can handle over 40,000 search queries every second on average, which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide. With such huge volume of search activities it is essential to make the search results of high quality in order to meet the users needs.

Information Retrieval (IR), usually used by academia in favor of its industrial counterpart search engine, is one of the most evolving fields and has drawn extensive attention in recent years. The primary goal of IR research is to improve the effectiveness or the efficiency or both of the textual retrieval system. There are many related works dedicated to this line of research already [16, 17, 22, 9, 19, 12]. Like everything in the world the various IR researches have their own context. The context could be the experiment settings. It can also be the assumptions on which the ranking model relies.

The word "*Context*" is originally defined as "the set of circumstances or facts that surround a particular event, situation, etc." In IR research the context always plays an crucial role and it is the fundamental basis of some very important IR research efforts. In our work, we pick three distinct domains to show the existence and the impact of the context.

[1] http://www.internetlivestats.com/google-search-statistics/

1.1 Unified IR Evaluation System

The first domain is related to the evaluation of IR system. There are many aspects that an IR system can be evaluated. In our work we focus on evaluating the effectiveness of the system. Many different techniques can be applied to address the effectiveness of the system. For example, Natural Language Processing (NLP) techniques [18, 14]. Topic Modeling [2, 11]. But most of the previous works target on the simple yet effective ranking models which usually applied to the document index. For a typical IR evaluation system the ideal case is to have a unified testing environment which is responsible for everything related to the evaluation process except the ranking model part. That said, everything including pre-processing and indexing the documents, ranking the documents, evaluating the results, the choice of evaluation metrics and interpretation of the performance should under the same setting if one's purpose is purely compare the effectiveness of different ranking models. Here the unified testing environment can be regarded as the context of the evaluation process.

The context here is the basis of any kind of comparison between ranking models. Without the unified testing environment people cannot make sound claim about their proposed models. Unfortunately, there is no such environment for the IR community. People continuously report different performances on the same baseline model [21] and this casts doubt on the real effectiveness of the proposed models.

In our work, two systems, namely VIRLab [8] and RISE [21] are proposed to specifically address the problem. The uniqueness and the advantage of these two system is that they offer centralized and controlled IR evaluation systems which facilitate the fair comparison of retrieval models. The systems are the instantiation and expansion of Privacy Preserving Evaluation (PPE)[8] and Evaluation as a Service (EaaS)[15]. With the help of these systems (especially RISE) we are able to conduct a comprehensive reproducibility study for information retrieval models. In particular, we implement and evaluate more than 20 basic retrieval functions over 16 standard TREC collections. Experimental results allow us to make a few interesting observations. We first compare the evaluation results with those reported in the original papers, and find that the

performance differences between the reproduced results and the original ones are small for majority of the retrieval functions. Among all the implemented functions, only one of them consistently generates worse performance than the one reported in the original paper. Moreover, we report the retrieval performance of all the implemented retrieval functions over all the 16 TREC collections including recently released ClueWeb sets. To the best of our knowledge, this is the first time of reporting such a large scale comparison of IR retrieval models. Such a comparison can be used as the performance references of the selected models.

1.2 Boundary Theory of Bag-of-Terms Models

Classic IR ranking models [16, 17, 22, 1, 9, 13, 10] are mainly based on bag-of-terms document representation assumption and they mainly consist of basic signals (statistics) such as Term Frequency (TF), Inverted Document Frequency (IDF), Document Length Normalization (DLN) and other collection statistics [7]. For this domain we can view the bag-of-terms assumption and the commonly used statistics as the context of the ranking models since they are the theoretical foundation of these retrieval functions. Under this context our question is whether we have reached the performance upper bound for retrieval functions using only basic ranking signals. If so, what is the upper bound performance? If not, how can we do better?

To find the performance upper bound is quite challenging: although most of the IR ranking models deal with basic signals, how they combine the signals to compute the relevance scores are quite diverse due to different implementations of IR heuristics [7]. This kind of variants makes it difficult to generalize the analysis. Moreover, typically there are one or more free parameters in the ranking models which can be tuned via the training collections. These free parameters make the analysis more complicated. In our work, we simplify the problem and just focus on single-term queries and study how to estimate the performance bound for retrieval functions utilizing only basic ranking signals. With only one term in a query, many retrieval functions can be greatly simplified. For example, Okapi BM25 and Pivoted normalization functions

have different implementations for the IDF part, but this part can be omitted in the functions for single-term queries because it would not affect the ranking of search results. All the simplified functions can then be generalized to a general function form for single-term queries. As a result, the problem of finding the upper bound of retrieval function utilizing basic ranking signals becomes that of finding the optimal performance of the generalized retrieval function. We propose to use cost/gain analysis to solve the problem [3, 4, 6]. As the estimated performance upper bound of simplified/generalized model is in general better than the existing ranking models, our finding provides the practical foundation of the potentially more effective ranking models for single term queries.

1.3 Contextual Suggestion

The task of contextual suggestion is to recommend interesting venues to the users based on contextual information such as geographic location, temporal information and user's activity history. Context again, as the name of the problem shows, highlights this direction of research effort.

There are two necessary steps to tackle the contextual suggestion problem: (1) identify a set of candidates that satisfy the contextual requirements, e.g., places of interest that are close to a target place; (2) rank the candidates with respect to the user interest. In our work we focus on the second problem assuming the first requirement has already been fulfilled. User profiling is the key component to effectively rank candidate places with respect to a user's information need and the question for us is how to effectively model the user profile? We first propose to leverage the category and description information about the places in user's activity history to construct user profiles [20]. The advantage of such approach is the ease of computation and the satisfactory results [5]. We further find that using category or description to build a user profile is not enough: category of places is too general to capture a user's underlying needs; while the text description of a place is too specific to be generalized to other places. In another study we leverage opinion, i.e. opinion ratings and the associated

text reviews, to construct an opinionated user profile. By doing like this we aim to explain "why the user likes or dislikes thesuggestion" instead of simply recording "what places the user liked or dislike" in the search history. The problem of this approach is that on-line opinions are notoriously skewed as only very small number of people post their opinions. To address this data sparsity challenge we propose to also include the opinions from similar users as the current user to construct the profile of current user. The assumption here is that users with similar ratings have the similar reasons of giving the rating. By modeling the candidate places in the similar fashion the similarity between user profile and candidates profile is used to rank the candidates. We tried different representations of the text reviews when modeling the profiles. We further apply Learning-to-Rank (LTR) method to the similarity scores for the ranking method. Experiment results on TREC collections and a self-crawled Yelp collection validate the effectiveness of our method.

Previous studies in IR rarely separated the context apart from other components of IR research problem. However, we argue that the context of different IR systems or components should be carefully treated and extensively studied as the impact of the context is big enough to question the foundation of some IR studies. For example, the evaluation streamline is one of the key components of IR system where different ranking approaches can be easily compared. Moreover, there are many other web applications which are highly related to the IR system. One of such domain is recommendation system where researchers tried their best to incorporate IR techniques with this area hoping to satisfy users different needs.

Apparently, the context of every line of research effort in IR should be standardized so that the full picture could be seen clearly. However, there is few literatures dedicated to elaborate on the context of these research efforts.

Without clearly defining of the context of each line of research we might get dim results and this is not a good sign for the whole IR community. Here we show three domains that extensively studied in this dissertation.

The primary goal of IR research is to effectively address user's information needs

such as search via text queries or recommendation based on historical activities.

Opinions are also used to generate personalized and high quality summaries of the suggestions.

BIBLIOGRAPHY

[1] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] Christopher J.C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, June 2010.

[4] C.J.C. Burges, R. Ragno, and Q.V. Le. Learning to rank with non-smooth cost functions. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, January 2007.

[5] Adriel Dean-Hall, Charles Clarke, Jaap Kamps, Paul Thomas, and Ellen Voorhees. Overview of the trec 2012 contextual suggestion track. In *Proceedings of TREC'12*, 2012.

[6] Pinar Donmez, Krysta M. Svore, and Christoper J.C. Burges. On the local optimality of lambdarank. In *SIGIR*. Association for Computing Machinery, Inc., July 2009.

[7] Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 49–56, New York, NY, USA, 2004. ACM.

[8] Hui Fang, Hao Wu, Peilin Yang, and ChengXiang Zhai. Virlab: A web-based virtual lab for learning and studying information retrieval models. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1249–1250, New York, NY, USA, 2014. ACM.

[9] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 480–487, New York, NY, USA, 2005. ACM.

[10] Ben He and Iadh Ounis. A study of the dirichlet priors for term frequency normalisation. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 465–471, New York, NY, USA, 2005. ACM.

[11] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[12] Xitong Liu and Hui Fang. Latent entity space: a novel retrieval approach for entity-bearing queries. *Information Retrieval Journal*, 18(6):473–503, 2015.

[13] Yuanhua Lv and ChengXiang Zhai. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 7–16, New York, NY, USA, 2011. ACM.

[14] Rada F. Mihalcea and Dragomir R. Radev. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge University Press, New York, NY, USA, 1st edition, 2011.

[15] Jinfeng Rao, Jimmy Lin, and Miles Efron. Reproducible experiments on lexical and temporal feedback for tweet search. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*, pages 755–767. Springer International Publishing, 2015.

[16] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. pages 109–126, 1996.

[17] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 21–29, New York, NY, USA, 1996. ACM.

[18] Ellen M. Voorhees. Natural language processing and information retrieval. In *Information Extraction: Towards Scalable, Adaptable Systems*, pages 32–48, London, UK, UK, 1999. Springer-Verlag.

[19] Hao Wu and Hui Fang. An incremental approach to efficient pseudo-relevance feedback. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 553–562, New York, NY, USA, 2013. ACM.

[20] Peilin Yang and Hui Fang. An exploration of ranking-based strategy for contextual suggestion. In *Proceedings of TREC'12*, 2012.

[21] Peilin Yang and Hui Fang. A reproducibility study of information retrieval models. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 77–86, New York, NY, USA, 2016. ACM.

[22] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.