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

Chapter

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. For example, **Citation Here!!!!!** [1]. However, there is few literatures dedicated to elaborate on the context of these research efforts.

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. 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.

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

For example, classic IR ranking models are mainly based on bag-of-terms document representation assumption and they mainly consist of statistics such as Term Frequency (TF), Inverted Document Frequency (IDF), Document Length Normalization (DLN) and other collection statistics. Here the bag-of-terms assumption and the commonly used statistics are the context of the ranking models. Another example comes from the IR evaluation. For evaluation of IR system the ideal case is we use a unified testing environment to assess the ranking models. The evaluation process is then purely based on the algorithms since the unified testing environment takes care of the possible processing stages, e.g. prepare the data and standardize the results. Here the ideal unified testing environment is the context of the evaluation process. For the third example we choose one of the IR recommendation system which is called contextual suggestion. For contextual suggestion the problem 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. 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.

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] 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.