Chapter 3
From Tf-Idf to Learning-to-Rank: An Overview

Muhammad Ibrahim
Monash University, Australia

Manzur Murshed
Federation University Australia

ABSTRACT

Ranking a set of documents based on their relevances with respect to a given query is a central problem of information retrieval (IR). Traditionally people have been using unsupervised scoring methods like tf-idf, BM25, Language Model etc., but recently supervised machine learning framework is being used successfully to learn a ranking function, which is called learning-to-rank (LtR) problem. There are a few surveys on LtR in the literature; but these reviews provide very little assistance to someone who, before delving into technical details of different algorithms, wants to have a broad understanding of LtR systems and its evolution from and relation to the traditional IR methods. This chapter tries to address this gap in the literature. Mainly the following aspects are discussed: the fundamental concepts of IR, the motivation behind LtR, the evolution of LtR from and its relation to the traditional methods, the relationship between LtR and other supervised machine learning tasks, the general issues pertaining to an LtR algorithm, and the theory of LtR.

1. INTRODUCTION

In the last few decades, there has been an overwhelming increase in the volume of digital data due to the proliferation of information and communications technology. Getting the required information from this vast ocean of data has eventually become so formidable that people started using machines from late 1980's[1] to get assistance, thereby giving rise to information retrieval (IR) systems (i.e., search engines). In general, the task of an IR system is to return a ranked list of 'items' to the users in response to specific information need. This task appears in many domains such as document ranking, recommender system, automatic question answering, automatic text summarization, online

DOI: 10.4018/978-1-4666-8833-9.ch003

advertising, sentiment analysis, web personalization, and so on. In fact, any task which presents the user on-demand a list of items ordered by a utility function is a ranking task. In this chapter, we survey only the document ranking research works without any loss of generality, as most of the discussed techniques are applicable to other ranking domains as well. We use the terms IR ranking and document ranking interchangeably throughout the chapter.

1.1 Scope of the Chapter

Several standard books on IR (e.g., Manning, Raghavan & Schütze, 2008) are available in the literature. These books, however, do not cover the learning-to-rank (LtR) systems with appropriate emphasis, mainly because these systems have emerged as a promising IR direction only a few years ago. There are a few survey papers on LtR; some of these are more focused on detailed discussion of the technical aspects of the LtR algorithms (Li, 2011; Liu, 2011), while some others are too short (Phophalia, 2011; He, Wang, Zhong & Li, 2008). Therefore, these reviews provide very little assistance to someone who, before delving into technical details of different algorithms, wants to have a broad understanding of the LtR systems, and their evolution from and relation to the traditional IR methods.

This chapter is complementary to the existing few surveys in the sense that we focus on the evolution of the LtR systems from the conventional methods. It also elaborately discusses some aspects of LtR that have so far been less-emphasized. Specifically, our main goals are the following:

1. Familiarise the readers with the fundamental concepts of IR from scratch by emphasising on intuitive explanations.
2. Discuss the motivation behind LtR; how it has evolved from and relates to the traditional IR methods.
3. Show the relationship between LtR and other supervised machine learning tasks, namely, classification, regression, and ordinal regression.
4. Discuss the general issues pertaining to the LtR algorithms. That is, to give a big picture of the existing LtR algorithms before delving into technical details of individual algorithms. Some of these issues are: relationship between LtR and other machine learning tasks and developing taxonomy of LtR algorithms.
5. Relate the theory of LtR to various loss functions of existing LtR algorithms.

This chapter is not a comprehensive survey of all LtR algorithms—in fact, it is not feasible to discuss all of them in a single chapter as the number of research papers on LtR is more than a hundred, nor does it discuss the technical details of different algorithms.

1.2 Organization of the Chapter

The rest of the the chapter is organized as follows.

- Section 2 defines the ranking problem in IR.
- Section 3 explains why existing search engines are not sufficient.
- Section 4 describes the term *relevancy* with a simple but concrete example.
- Section 5 briefly discusses three popular IR models, namely, cosine similarity, BM25, and Language Model, with an emphasis on intuitive explanations.
- Section 6 argues why the traditional IR models need further improvement, and hence a remedy can be found by using the supervised machine learning framework. Here some very basic concepts of machine learning are presented. Then the strategies for preparing a training set of LtR is de-