Modeling Score Distributions

Anca Doloc-Mihu
University of Louisiana at Lafayette, USA

INTRODUCTION

The goal of a web-based retrieval system is to find data items that meet a user's request as fast and accurately as possible. Such a search engine finds items relevant to the user's query by scoring and ranking each item in the database. Swets (1963) proposed to model the distributions of these scores to find an optimal threshold for separating relevant from non-relevant items. Since then, researchers suggested several different score distribution models, which offer elegant solutions to improve the effectiveness and efficiency of different components of search systems.

Recent studies show that the method of modeling score distribution is beneficial to various applications, such as outlier detection algorithms (Gao & Tan, 2006), search engines (Manmatha, Feng, & Rath, 2001), information filtering (Zhang & Callan, 2001), distributed information retrieval (Baumgarten, 1999), video retrieval (Wilkins, Ferguson, & Smeaton, 2006), kernel type selection for image retrieval (Doloc-Mihu & Raghavan, 2006), and biometry (Ulery, Fellner, Hallinan, Hicklin, & Watson, 2006).

The advantage of the score distribution method is that it uses the statistical properties of the scores, and not their values, and therefore, the obtained estimation may generalize better to not seen items than an estimation obtained by using the score values (Arampatzis, Beney, Koster, & van der Weide, 2000). In this chapter, we present the score distribution modeling approach, and then, we briefly survey theoretical and empirical studies on the distribution models, followed by several of its applications.

BACKGROUND

The primary goal of information retrieval is to retrieve all the documents which are relevant to a user query, while retrieving as few non-relevant documents as possible (Baeza-Yates & Ribeiro-Neto, 1999). This is achieved by ranking the list of documents according to their relevance to the user's query. Since relevance is a subjective attribute, depending on the user's perception of the closeness between the user submitted query and the real query from her or his mind, building a better way to retrieve data is a challenge that needs to be addressed in a retrieval system.

In other words, a retrieval system aims at building the request (query) that best represents the user's information need. This optimal request is defined by using an explicit data-request matching (Rocchio, 1971) that should produce a ranking in which all relevant data are ranked higher than the non-relevant data. For the matching process, a retrieval system uses a retrieval function, which associates each data-query pair with a real number or score (the retrieval status value). Then, the retrieval system uses these scores to rank the list of data.

However, researchers (Swets, 1963; Arampatzis, Beney, Koster, & van der Weide, 2000; Manmatha, Feng, & Rath, 2001) raised the question of whether or not the statistical properties of these scores, displayed by the shape of their distribution, for a given query, can be used to model the data space or the retrieval process. As a result, they proposed and empirically investigated several models of the score distributions as solutions to improve the effectiveness and efficiency of the retrieval systems. The next section introduces the score distribution method.

MAIN FOCUS

The Score Distribution Method

The probability ranking principle (Robertson, 1977) states that a search system should rank output in order of probability of relevance. That is, the higher the score value of the document, the more relevant to the query is considered the document to be. In the binary relevance case, which is the case we are interested in, the ideal retrieval system associates scores to the relevant and non-relevant data such that the two groups are well

separated, and relevant data have higher scores than the non-relevant data. In practice, retrieval systems are not capable to completely separate the relevant from the non-relevant data, and therefore, there are non-relevant data with higher score values than those of some relevant data.

The score distribution method tries to find a good way to separate these two groups of data by using statistical properties of their scores. The method assumes that the relevant and non-relevant data form two separate groups, with each group being characterized by its own characteristics different from the other group. For each group, the method plots the corresponding score values within the group, and then, tries to find the shape of the curve generated by these scores. In fact, this curve is approximated with a distribution usually chosen via experimental results (the best fit from a set of known distributions, such as normal, exponential, Poisson, gamma, beta, Pareto). Once the two distributions are known (or modeled), they are used to improve the search system.

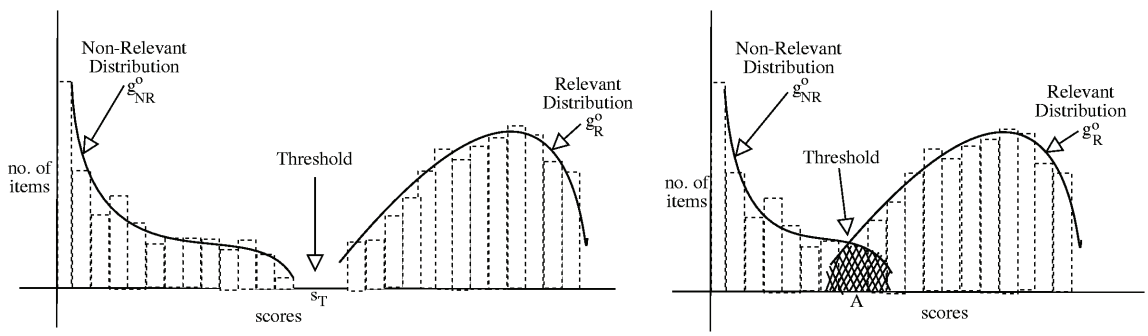
Figure 1 illustrates the score distribution method, (a) in the ideal case, when the relevant and non-relevant data are well separated by the retrieval system, and (b) in a real case, when there are non-relevant data with score values higher than those of some relevant data. The scores of non-relevant data are grouped toward the left side of the plot, and the scores of relevant data are grouped toward the right side of the plot. A curve shows the shape of the score distribution of each group (of relevant and non-relevant data, respectively). Note that, in this figure, the two curves (given as densities $g^{O}_{R}(s)$ and $g^{O}_{NR}(s)$) do not display any particular distribution; they represent the curves of some arbitrary distributions. Basically, the score distribution method consists in choosing the best possible shapes of the distributions of the two groups. Then, any relevant (non-relevant) data is assumed to follow its chosen relevant (non-relevant) distribution.

Ideally, the two distribution curves do not meet (Figure 1 (a)), but in reality, the two curves meet at some point. However, as shown in Figure 1 (b), there is a common region between the two score distribution curves (named A). This area is of most interest for researchers; it includes relevant data with score values very close (lower or not) to the score values of non-relevant data. Therefore, by finding a way to minimize it, one finds a way to approximately separate the two data. Another solution is to find a threshold that separates optimally the relevant data from non-relevant ones.

The advantage of the score distribution method is that it uses the statistical properties of the scores (the shape of their distribution) and not their values, which conducts to an estimation of the threshold or the area A (Figure 1 (b)) that may generalize better to not seen data than an estimation method, which uses the score values (Arampatzis, Beney, Koster, & van der Weide, 2000).

We presented the method in the case that the entire data from collection is used. However, for efficiency reason, in practice, researchers prefer to return to user only the top most relevant N data. In this case, as Zhang and Callan (2001) noticed, the method is

Figure 1. Score distributions for relevant and non-relevant data



(a) Ideal case, at which a retrieval system aims, with a clear separation between the relevant and non-relevant data.

(b) Real case, which shows a common region for scores of the relevant and non-relevant data.