Chapter 83
Mining Statistically Significant Substrings based on the Chi-Square Measure

Sourav Dutta
IBM Research Lab, India

Arnab Bhattacharya
Indian Institute of Technology Kanpur, India

ABSTRACT

With the tremendous expansion of reservoirs of sequence data stored worldwide, efficient mining of large string databases in various domains including intrusion detection systems, player statistics, texts, and proteins, has emerged as a practical challenge. Searching for an unusual pattern within long strings of data is one of the foremost requirements for many diverse applications. Given a string, the problem is to identify the substrings that differ the most from the expected or normal behavior, i.e., the substrings that are statistically significant (or, in other words, less likely to occur due to chance alone). We first survey and analyze the different statistical measures available to meet this end. Next, we argue that the most appropriate metric is the chi-square measure. Finally, we discuss different approaches and algorithms proposed for retrieving the top-k substrings with the largest chi-square measure.

INTRODUCTION

Detection or identification of statistically significant sequences or mining interesting patterns from a given string has lately emerged as an important area of study (Denise et al., 2001; Ye & Chen, 2001). In such applications, we are given an input string composed of symbols from an alphabet set with a probability distribution defining the chance of occurrence of each symbol, and the aim is to find those portions of the string that deviate most from their expected nature, and are thus potent sources of hidden pattern and information. Such solutions come handy in automated monitoring systems, such as in a cluster of sensors sensing the ambient temperature for possible fire alert, or

DOI: 10.4018/978-1-4666-3604-0.ch083

a network server sniffing the network for intrusion detection. Also, text analysis of blogs, stock market trend deciphering, detection of protein mutation and the identification of good and bad career patches of a sports icon can be few of the target applications. It is such diverse utility that makes the study and development of this field challenging and necessary.

STATISTICAL MODELS AND TOOLS

Establishing a relationship of the empirical or observed results of an experiment to factors affecting the system or to pure chance calls for various statistical models and measures. In such scenarios, an observation is deemed statistically significant if its presence cannot be attributed to randomness alone. The literature hosts a number of statistical models to capture the uniqueness of such observations such as *p-value* and *z-score*. In the next few sections, we discuss different important statistical tools that are used for this purpose.

Before venturing forward, we provide a formal definition of the problem.

Problem 1: Given a string S of length l comprising symbols from the alphabet set Σ of cardinality m, and with a given probability distribution P modeling the chance of occurrence of each symbol in Σ, the problem is to efficiently identify and extract the top-k substrings that exhibit the largest deviation from the expected nature, i.e., the substrings that are most statistically significant.

It is this measure of deviation of a sequence that we will capture by using various statistical models. In the remainder of the chapter, we interchangeably use the term string with sequence and substring with subsequence.

Hypothesis Testing and P-Value

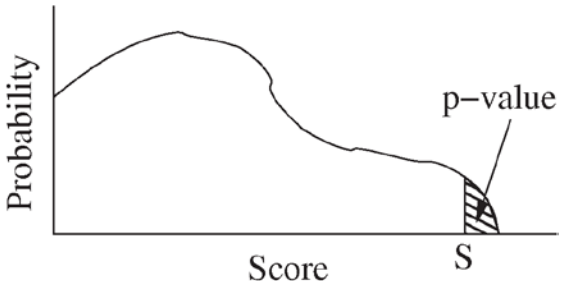
Given an observation sample X (in this case a substring), with an associated score of $S(X)$, the *p-value* of X is defined as the probability of obtaining a random sample with score $S(X)$ or greater under the same probability model (Bejerano et al., 2004; Regnier & Vandenbogaert, 2006). For each such observation, we test the null hypothesis H_0 that the substring is drawn from the given probability model P against the alternate hypothesis H_1 that the subsequence is not drawn from the same probability distribution. The p-value measures the chance of rejecting the null hypothesis; in other words, the less the p-value, the less likely it is that the null hypothesis is true.

Figure 1 shows an example. For a particular score S, the shaded area represents the chance of having a sample with a score greater than the one under consideration. In other words, the p-value is the value of the cumulative density function (cdf) measured at S subtracted from the total probability, i.e.,

pvalue(S)=1-cdf(S).

If the probability density function (pdf) of the scores is known, it is relatively simpler to compute the p-value of a particular score using the above formula. However, in most real situations, the pdf is hard to estimate or can be non-parametric. The accurate computation of the p-value then needs all the possible outcomes to be listed, their scores

Figure 1. Computing the p-value of X with score S


Probability
p−value
Score
S