Cycle Detection and Correction

Amihood Amir[1,2,⋆], Estrella Eisenberg[1,⋆⋆], Avivit Levy[3,4], Ely Porat[1], and Natalie Shapira[1]

[1] Department of Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel
{amir,porately,davidan1}@cs.biu.ac.il
[2] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
[3] Department of Software Engineering, Shenkar College,
12 Anna Frank, Ramat-Gan, Israel
avivitlevy@shenkar.ac.il
[4] CRI, Haifa University, Mount Carmel, Haifa 31905, Israel

Abstract. Assume that a natural cyclic phenomenon has been measured, but the data is corrupted by errors. The type of corruption is application-dependent and may be caused by measurements errors, or natural features of the phenomenon. This paper studies the problem of recovering the correct cycle from data corrupted by various error models, formally defined as the *period recovery problem*. Specifically, we define a metric property which we call *pseudo-locality* and study the period recovery problem under pseudo-local metrics. Examples of pseudo-local metrics are the Hamming distance, the swap distance, and the interchange (or Cayley) distance. We show that for pseudo-local metrics, periodicity is a powerful property allowing *detecting* the original cycle and *correcting* the data, under suitable conditions. Some surprising features of our algorithm are that we can *efficiently* identify the period in the corrupted data, up to a number of possibilities logarithmic in the length of the data string, even for metrics whose calculation is $\mathcal{NP}$*-hard*. For the Hamming metric we can reconstruct the corrupted data in near linear time even for unbounded alphabets. This result is achieved using the property of *separation* in the self-convolution vector and Reed-Solomon codes. Finally, we employ our techniques beyond the scope of pseudo-local metrics and give a recovery algorithm for the non pseudo-local Levenshtein edit metric.

1 Introduction

Cyclic phenomena are ubiquitous in nature, from Astronomy, Geology, Earth Science, Oceanography, and Meteorology, to Biological Systems, the Genome, Economics, and more. Part of the scientific process is understanding and explaining these phenomena. The first step is identifying these cyclic occurrences.

Assume, then, that an instrument is making measurements at fixed intervals. When the stream of measurements is analyzed, it is necessary to decide whether these measurements represent a cyclic phenomenon. The "cleanest" version of

⋆ Partially supported by NSF grant CCR-09-04581 and ISF grant 347/09.
⋆⋆ Supported by a Bar-Ilan University President Fellowship.

S. Abramsky et al. (Eds.): ICALP 2010, Part I, LNCS 6198, pp. 43–54, 2010.

this question is whether the string of measurements is *periodic*. Periodicity is one of the most important properties of a string and plays a key role in data analysis. As such, it has been extensively studied over the years [16] and linear time algorithms for exploring the periodic nature of a string were suggested (e.g. [12]). Multidimensional periodicity [4,14,18] and periodicity in parameterized strings [6] was also explored.

However, realistic data may contain errors. Such errors may be caused by the process of gathering the data which might be prone to transient errors. Moreover, errors can also be an inherent part of the data because the periodic nature of the data represented by the string may be inexact. Nevertheless, it is still valuable to detect and utilize the underlying cycle. Assume, then, that, in reality, there is an underlying periodic string, which had been corrupted. Our task is to discover the original uncorrupted string.

This seems like an impossible task. To our knowledge, reconstruction or correction of data is generally not possible from raw natural data. The field of Error Correcting Codes is based on the premise that the original data is not the transmitted data. Rather, it is converted to another type of data with features that allow correction under the appropriate assumptions. Without this conversion, errors on the raw data may render it totally uncorrectable. A simple example is the following: consider the string *aaaaa aaaaa aaaaa aaaab*. This may be the string *aaaaa aaaaa aaaaa aaaaa* with one error at the end (an a was replaced by a b), or the string *aaaaa aaaab aaaaa aaaab* with one error at the 10th symbol (a b was replaced by an a). How can one tell which error it was?

In this paper we show that, surprisingly, data periodicity acts as a feature to aid the data correction under various error models. The simplest natural error model is substitution errors. It generally models the case where the errors may be transient errors due to transmission noise or equipment insensitivity.

Of course, too many errors can completely change the data, making it impossible to identify the original data and reconstruct the original cycle. On the other hand, it is intuitive that few errors should still preserve the periodic nature of the original string. The scientific process assumes a great amount of confidence in the measurements of natural phenomena, otherwise most advances in the natural sciences are meaningless. Thus, it is natural to assume that the measurements we receive are, by and large, accurate. Therefore, it is reasonable to assume that we are presented with data that is faithful to the original without too many corruptions. Formally, the problem is defined as follows:

The Period Recovery Problem. Let S be n-long string with period P. Given S', which is S possibly corrupted by at most k errors under a metric d, return P.

The term "recovering" is, in a sense, approximating the original period, because it may be impossible to distinguish the original period from other false candidates. The "approximation" of the original period means identifying a small set of candidates that is guaranteed to include the original period. We are able to provide such a set of size $O(\log n)$.