Data mining of website meta-data

Nils Lorenz Kjær Lück and Anders Høst Kjærgaard
{nikj, ahkj}@itu.dk

IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark

Abstract. This report describes novel approach to data mining of websites. Being able to find patterns and statistics in the diverse landscape of the web, is interesting for anyone related to online businesses. Based on a novel business concept developed by company, Statistico, residing in Aarhus, Denmark. We present the results of data mining project, seeking to find answers in the kind of meta-data that the company has focus on. We show that a significant amount of information can bed gathered on most websites, from data publicly available. We also give examples of what patterns can be found in the meta-data in question, and argue that some indications and predictions might be possible to find . The implications of performing this kind of data mining could potentially have impact on web businesses, as we can arm these with extra tools and information to out perform competitors.

Keywords: Data mining, Website

Is this true? Write what our real and most significant findings are

Abstract is a bit too long. Elements are probably what they should be, but write it more concisely.

1 Introduction

Refer to Web shops and the large and extremely competitive market of online marketing. Important to rank high on google and Alexa. Statistico general interest in these figures, and have a novel idea of scraping all web data, and make meta data searchable. What does their data mean? Patterns in server, CMS, HTML version. Stats are of general interest, and there are actually not much clean data and overview of the landscape of websites and their meta data.

1.1 Data set

Introduce the size of Statistico and the problem of delivering the data. Present right away that we do not have time to run in on their data set, but that we have collected a subset, and that our setup would work on the big set, but would take 17 days to run.

Motivation Why we think this is interesting. We know that the data set will be difficult to derive anything from, but we find that there is actually a real world scenario where our results could be useful, if the project is worked on in the future.

Overview Present sections and their content.

2 Background

In this section we present the problem statement, which addresses question relevant to Statistico, combined with our own interest in experimenting with certain data mining techniques. We also provide a short introduction to Alexa, Solr, and some of the technologies used in the solution

2.1 Problem Statement

Based on the interests and data of Statistico introduced in secton 1, we want to find out how well we can extract general statistics on websites in the .dk domain. We want to be able to give an overview of the distribution of CMS, server software, and SEO related figures. SEO related data could be related to link counts or presence of certain HTML tags. In addition, we want to investigate if it is possible to derive interesting correlations or patterns in the meta-data.With Aprior we can try and detect frequent patterns in the meta-data, and we will try to answer if there is anything interesting to say about the frequent patterns that we can get from the scraped data. A very important measure for any website is its Alexa-rank or page-rank on Google. We will see if we can predict these values based on the limited data we get. It seems unlikely, or at least not intuitive, that we should be able to derive these numbers from a data mining process that is essentially a simplification of what Alexa and Google's crawlers do, but non the less, the results could be interesting in that it could indicate a small set of parameters to be important for getting a high rank.

Site	Bottom 5
border-shop.dk	3
dotseo.dk	3
c-lager.dk	3
kreacom.dk	3
grafical.dk	3

Site	Top 5
www	5
www.medk	5
www.aiu.dk	5
www.mcdonalds.dk	5
www.mcb.dk	5

2.2 Meta data

Present meta data. What are we collecting and why. Justify these with either interest in SEO, or web statistics in general.

2.3 Alexa

Describe Alexa, grep the web, Statistico's ́big brother; what data does Alexa provide.

2.4 Solr

Statistico uses Solr. How it is used - ie their idea. Why we probably do not want to go that way. Solr is a search engine - schema layer just adds extra layer of complexity.

2.5 Scraping

HTML parser in Python is crap. Explain issue - HTML is extremely noisy. Present beautiful soup with lxml, our contribution to others trying to do the same.

3 Solution

We now present our solution, which primarily deals with collection and preprocessing the data. We have implemented a small preprocessing library that uses a plugin architecture that allows us to add, remove, and alter parser/scanners.

3.1 Preprocessing

Class-like diagram showing analyser, scanners, website, scraper, calls to alexa.

3.2 Weka

How is data fed to weka. Problems?

4 Evaluation

Diagrams and numbers. Evaluate and reflect upon all results. Can we say anything?

I imagine this is our prime section.

5 Conclusion

Conclusion, short and in 'summary' turns, highlighting most interesting findings.

A Appendix A

Sample appendix