Using Supervised Machine Learning to Identify Swiss Companies' Websites

Zeno Bardelli

Master thesis in Computer Science

Finding companies website is a fundamental task for many applications ranging from building a database of business to automatic collection of data directly from the websites. Yet, automatically identifying the website of a company is challenging. Existing methods rely heavily on the first result shown by search engines. However, the first result presented by search engines does not necessarily correspond to the company website. In some cases, the first result is the website of a company with a similar name, in other cases it is a page on a yellow pages website. The solution proposed in this work is to identify a company's website given its properties and the results of Google Search. We have conducted extensive experiments with existing machine learning classifiers on a data set of Swiss companies. We have studied multiple methods to improve the performance of these classifiers, including hyperparameters tuning, features selection and post processing. Experimental results on identifying Swiss companies website show that our approach improves the results obtained by a Google search by 26.60% in terms of F1 score, from 62.58% to 89.18%.

Prof. Philippe Cudré-Mauroux