SwissFinder: Identifying Swiss Websites from Unstructured Content

Zeno Bardelli, Ines Arous, Philippe Cudré-Mauroux
U. of Fribourg—Switzerland
{zeno.bardelli, ines.arous, pcm}@unifr.ch

Ljiljana Dolamic
armasuisse—Switzerland
Ljiljana.Dolamic@armasuisse.ch

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378259

***Abstract*—Finding companies' websites is important when building business databases. However, automatically finding a company's website based on its name or its official entry in a registry is challenging, as companies often have similar names, acronyms, or descriptions. In this context, we built a system to evaluate different features and classifiers to automatically identify a company's website from unstructured content.**

***Index Terms*—Web searching and information discovery**

I. Introduction

Companies often search for potential collaboration when they need a particular type of good or service. They usually proceed by looking through the official register, which contains a vague description of all companies' activities. It is very frustrating for companies to contact or, worse, travel to a listed business location only to find out their service does not precisely match their needs.

Armasuisse Science and Technology, the R&D agency of the Swiss Armed Forces, is developing a *Business Collaboration* platform to allow companies to reach out to each other and foster collaboration. One of the main features of the platform is to leverage the companies websites, which usually provide *detailed* information on the companies and their activities. However, automatically finding a company's website based on its name only or based on minimal information from its official listing is very challenging, as company names are often short and ambiguous. Existing methods mainly rely on the first result shown by search engines, which can lead to a yellow page website or to an entity with a similar name.

In a collaboration between armasuisse and the University of Fribourg, we tackle this problem and propose to identify companies' websites based on minimal information using statistical machine learning models. Our resulting system, SwissFinder, achieves a 88% F1-score by leveraging new features for website identification.

II. Existing Services

Several products and APIs have been developed to identify companies' websites given their name. For example, *Name to Domain API* was developed by Clearbit [1]. Their approach consists of matching the company name with existing domains and returning the one with the most traffic. A similar API was developed by Phantombuster [2] and is called *Domain Name Finder*. This API takes the first result that appears by querying a company name on numerous search engines. Other platforms rely on a unique search engine such as the Blockspring [3] platform, which relies solely on Bing Search. Some solutions are dedicated to enriching company data such as the one developed by Powrbot [4]. It extracts not only the website but also the location and revenue given a company name.

Despite their high performance on finding the websites of established companies, these products fail to accurately identify local companies in Switzerland as they tend to consider only the first result shown by search engines, which often lead to a yellow page. SwissFinder tackles this problem by leveraging statistical machine learning models and using various features extracted from the top-10 results on Google.

III. Method

In this section, we first describe the dataset used in our experiments then present the SwissFinder algorithm.

A. Data Collection

Armasuisse collected a dataset named *Swiss-Search*. It consists of the top-10 results on Google of 48k companies' names, which results in 480k entries, among which 14% are labeled positive, i.e., the Google result matches the company's website. For each entry in *Swiss-Search*, we identified a set of 22 key features that include the rank of the search result, the title and the location indicated on the website for our matching problem.

B. Notations

We denote the set of company names as $\mathcal{N}$. Each company $n \in \mathcal{N}$ has a website w. We denote the set of all companies websites as $\mathcal{W}$. The subset of companies with known websites is denoted as $\mathcal{N}_L$ and we use $\mathcal{W}_L$ for their corresponding websites.

C. Algorithm

The complete workflow of SwissFinder is described in Algorithm 1. First, we query the company name in *local.ch*, an online registry for businesses in Switzerland (row 3). The result of the query is a Boolean variable r that indicates if the company's website is listed in the registry. In such case (i.e., r is true), we crawl and save it (row 4-5). Otherwise, we use a set of statistical classifiers trained on the *Swiss-Search* dataset to automatically identify the missing website (row 6-8). Finally, we return the identified website (row 9).

Authorized licensed use limited to: McGill University. Downloaded on November 28,2022 at 15:45:45 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: SwissFinder

```
Input  : Company names $\mathcal{N}$, $\mathcal{N}_L$ and $\mathcal{W}_L$
Output: Company websites $\mathcal{W}$
1 $\mathcal{W} = []$;
2 for $n \in \mathcal{N}$ do
3     $r$ = query $n$ in local.ch;
4     if $r == True$ then
5         $w$ = GetWebsite($n$);
6     else
7         classifier = train a classifier using $\mathcal{N}_L$ and $\mathcal{W}_L$;
8         $w$ = predict the website of $n$ using classifier;
9     $\mathcal{W} = [\mathcal{W}, w]$;
```

The classifiers we considered in row 7 of the Algorithm are: 1) a Random Forest Classifier (RFC), which constructs a set of decision trees and outputs a combination of their results; 2) a Support Vector Classification (SVC), which constructs a hyperplane in a high-dimensional space used for classification; 3) a Logistic Regression (LR) classifier, which estimates the probability of a data instance belonging to a class; and 4) a Multilayer Perceptron (MLP) that assigns weights to input features and maps the weighted inputs to a class. Each classifier is trained on features from a training set, tuned on a validation set, and tested on a test set.

IV. Experiments

This section presents experimental results evaluating the classifiers performance and the features importance in identifying the website of a company.

A. *Varying the Supervision Degree*

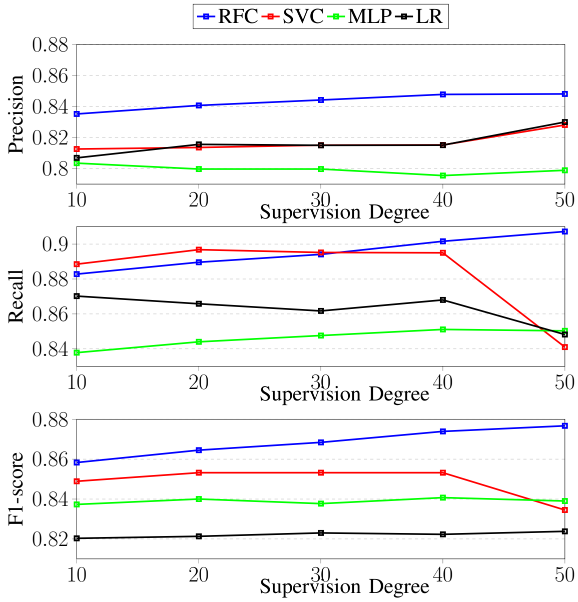


Fig. 1: Performance with varying supervision degree.

In what follows, we study the impact of the supervision degree on the performance of all classifiers to determine the minimum amount of ground truth needed. We split our datasets by s_{deg} where we vary s_{deg} between 10% and 50%, where $s_{deg} = 50\%$ means that we use 50% of the ground truth labels for training. The results are shown in Figure 1. We observe that all methods performance improve when increasing the size of the training set, in particular, RFC achieves the best performance for $s_{deg} = 50\%$. We also observe that SVC performance drops for $s_{deg} = 50\%$ while both MLP and LR have an overall stable performance when varying the supervision degree.

B. *Feature Importance*

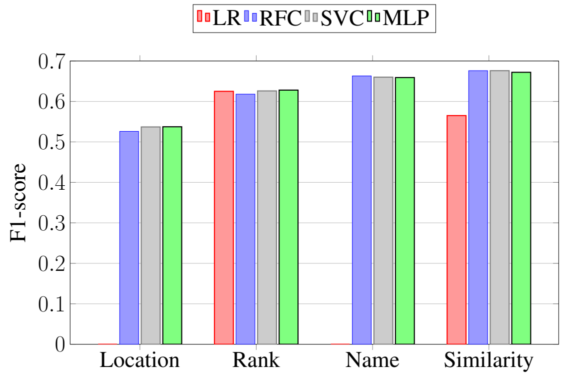


Fig. 2: Classifiers performance using a single feature

We compare each feature's importance by measuring the classifiers' performance in terms of F1-score with the selected feature. We find four features that are considered important by all classifiers. These features are: 1) *location*: indicates whether the company's location declared in the trade register appears on the web page. 2) *rank*: the rank of the Google search result. 3) *name*: indicates whether the company's name appears on the web page. 4) *similarity*: the string similarity distance between the Google result domain and the company name using the Jaccard metric. We omit the ranking of all 22 features and show the results for the four most important ones in Figure 2. These results confirm that features comparing the official information from the trade register (e.g., *location* and *name*) with the website content are the most valuable ones in identifying a company's URL and are even more important than the ranking of the webpage in a Google search.

C. *Feature Correlation*

We also study the correlation between the four most important features identified in the previous section. Results are shown in Figure 3. We find that the ranking of a website in a search result has a negative correlation with other features. This result was expected as the *rank* is a website property while all other features compares the information available on a website with the official documentation. We also find that *name* and *location* have a strong positive correlation as both the company name and its location tend to appear together on a website.

Authorized licensed use limited to: McGill University. Downloaded on November 28,2022 at 15:45:45 UTC from IEEE Xplore. Restrictions apply.

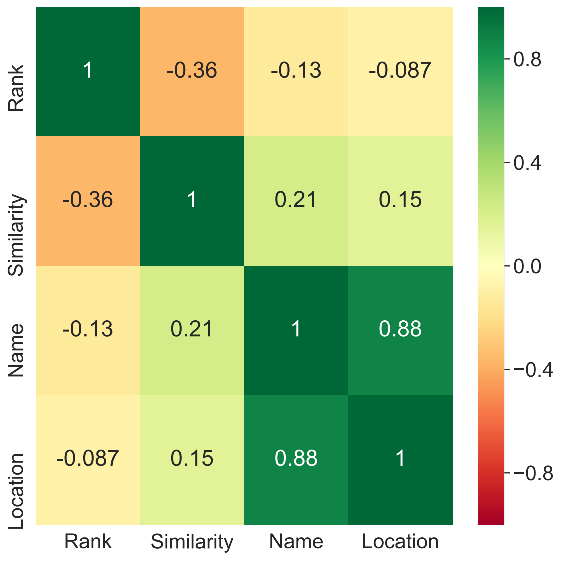


Fig. 3: Feature correlation matrix

D. *Ablation Analysis*

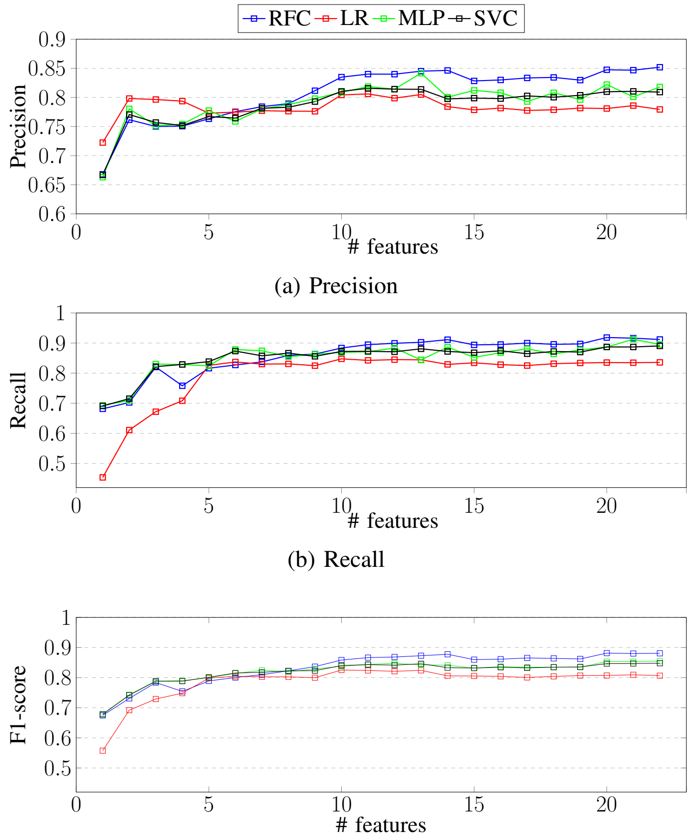
We use the ranking of features from Section IV-B and conduct an ablation analysis where we start by using only two of the most important features for all classifiers and incrementally add the others. The results are shown in Figure 4. We observe that while a probabilistic method such as Logistic Regression requires a small set of features to converge, other methods such as RFC require all features to reach optimal performance.

E. *Comparison with Existing Services*

In this section, we compare the performance of our method with the services described in Section II. As most of these services allow a free trial for a small set of names, we select randomly from *Swiss-Search* a set of 100 company names with known websites. Results are shown in Table I. We observe that products that use a single search engine such as Blockspring provide the best results followed by machine learning based products such as Powrbot. Most importantly SwissFinder has the best performance as it is better at leveraging website features for identifying the correct company website.

Service	**Owner**	**#Websites**
Company Name to Domain API	Clearbit	7
Domain Name Finder	Phantombuster	14
Company Data Enrichment	Powrbot	80
Company URL Lookup	Blockspring	84
SwissFinder	UNIFR & armasuisse	**85**

TABLE I: Comparison with existing services.



(a) Precision

(b) Recall

(c) F1-score

Fig. 4: Performance with varying number of features.

V. Discussion and Future Work

In this work, we built a system to identify websites from minimal input, and compared different features and classifiers in that context. We found that website features have different levels of importance. Most importantly, we observed that features comparing the company official information with the one found on a website are strong indicator of website identification.

In future work, we plan to use this observation to extract further features, e.g. by comparing the company activity in the trade register with its description on the website. We also plan to retrieve results from other search engines besides Google and leverage the ranking of websites on these engines to identify the correct website.

References

[1] “Company Name to Domain API,” https://clearbit.com/blog/company-name-to-domain-api/, Last accessed on 2020-08-09.

[2] “Domain Name Finder — Phantombuster,” https://phantombuster.com/automations/toolbox/3171/domain-name-finder, Last accessed on 2020-08-09.

[3] “Company URL Lookups using Bing Search - Blockspring,” https://open.blockspring.com/, Last accessed on 2020-08-09.

[4] “Find company information with our database search,” https://powrbot.com/company-search/, Last accessed on 2020-08-09.

Authorized licensed use limited to: McGill University. Downloaded on November 28,2022 at 15:45:45 UTC from IEEE Xplore. Restrictions apply.