UNIVERSITÀ DEGLI STUDI DI TRENTO

Department of Information Engineering and Computer Science

Bachelor degree in
Computer Science

Final Elaborate

Machine Learning Classifier: Matching Italian Companies with Facebook Pages and Users

A Random Forest Approach

Yannis Velegrakis
Supervisor

Michele Sordo
Graduand

Academic year 2015/2016

Acknowledgements

...thanks to...

Contents

Summary

Nowadays companies are getting more and more connected with potential customers with the help of social networks. These instruments are not only supposed to help people getting linked and chat together but also to give public figures and companies a way to show themselves and to get promoted. Social Networks directly provide services for advertisement: Facebook Ads for Facebook, Twitter Ads for Twitter etc. These services are developed to help people and organizations promoting themselves on that particular platform and usually aren't very useful as a platforms itselves to manage an entire lead generation campaign. Companies are supported only in the brand and products promotion, meaning that the ones that want to find potential customers are generally left alone. The lead generation process, the generation of consumer interest or inquiry into products or services of a business , requires company to have a list of potential customers and, most important, a way to connect with them.

The aim of this project is to show a practical solution to connect real companies with their social pages. Such solution is meant to be implemented in a bigger platform for lead generation that will help companies in the path of finding other businesses that may be interested in buying specific products. The solution shown connects Italian companies with their Facebook page in order to enrich companies' description on Atoka , a leading tool developed by SpazioDati that collects structured data and information of all the 6 million companies currently opened in Italy.

The first problem faced was to determine if companies really structure their facebook data in pages or if they use facebook users as pages. From a legal point of view it's against the Facebook Terms to use your personal account to represent something other than yourself (ex: your business), and you could permanently lose access to your account if you don't convert it to a Page. Taking care of the fact that Akota aims to give the more information it can about a company, team decided to do a first exploratory analysis to determine if italian companies really used one facebook pages. The results show that a lot of businesses still use facebook user as pages. The main reason found is that small companies can't afford to pay a single person to work on public relation, this cause the fact that the bigger companies are the more structured their facebook data was. [1]

In order to

1 Introduction

1.1 Introduction

One important part of companies' public relations is being connected with potential customers and make it easy to be found in the huge quantity of data available by potential clients. In a connected world, where many companies have a Facebook page (and many a website too), it's difficult to find somebody interested in buying your product or solution. As introduced before, social networks provide advertising services that are vertical to the particular platform (ex. Facebook, Twitter, etc.) but a marketing platform for lead generation service needs to link social networks' data with existing companies legally registered in the real world.

The paper analyze the process of connecting Italian companies with their Facebook pages or user profile. The aim of the project is to create a model that can automatically find if a Facebook page is the official (and real) page of a determined company. This job is very simple if it's done by a human but it can't be done by people because of time lacking: nobody could find the right match for more than 6 million companies in a reasonable amount of time. The only way to overcome human computational time and cost is to develop a prediction algorithm that can learn by itself to recognize an official page (or user) from a set of potential pages (and users).

The paper analyze only the subset of companies that has ATECO code n.56. ATECO n.56 is the code that collect companies that operate in the food sector . This limitation is caused by time reason: the classification time that the VPS will spend analyzing all the 6 million companies will probably be more than 20 days. Apart from time causes the process needs to be integrated with Spaziodati's information retrieval pipeline and the automation process made with Azkaban Flow is not described in this paper.

This problem was an important step in upgrading company information available on Atoka and gained priority at SpazioDati, making the company invest money and time to develop this project. This approach might also be extended to different information and usages in similar problems. The solution developed was born after discussion with SpazioDati's developers, that have big experience in comparable problems (like websites classification etc), taking care of the previous problem solution and hurdles found developing that.

The final solution is a machine learning algorithm application that runs after a score generating application that compute how similar is a page/user data collected to the company data already suited in Atoka Index. The machine learning algorithm chosen to classify the data in this project is Random Forest Classifier and the solution will explain why, in this situation, a random forest approach is better than other classification algorithms.

2 Motivation

2.1 Motivation

Before the project, only 1.62% of Italian companies had a Facebook Page/User associated in Atoka Index. The previous knowledge about facebook pages/user was collected analyzing the content of companies' official websites but was clearly not enough for a business product that aims to connect companies to new business customers.

The subset chosen is interesting to study because many restaurant, pubs and companies involved in the food sector:

- have a Facebook Page, or user created erroneously instead of a business page (now Facebook has built a tool that migrates a business user account to a formal Facebook page but not all the companies have used it so far).
- still don't have a website: this is a really fundamental task to improve contacts in Atoka and a great reason to find how many of them really have a Facebook Page.

Another interesting part of the problem is how to sample a training dataset for the classifier. In fact, as said before, there are only few companies with ATECO n.56 connected to a website and this implies that less of them have a Facebook page/user connected. The number of matched Facebook pages is 6,355, over a dataset of 362,154 companies with ATECO n.56, less than 2% of them.

The result that this project will produce will be used to upgrade Atoka index data and, in consequence, data available on the online Atoka lead generation platform.

3 Problem Statement

3.1 Problem Statement

Atoka Index stores information about Italian companies, including contacts like emails, official website and official Facebook page. Only the 1.62% of the companies had a linked Facebook Page or user before the project. The aim is to fill more companies with Facebook social data. This topic was chosen

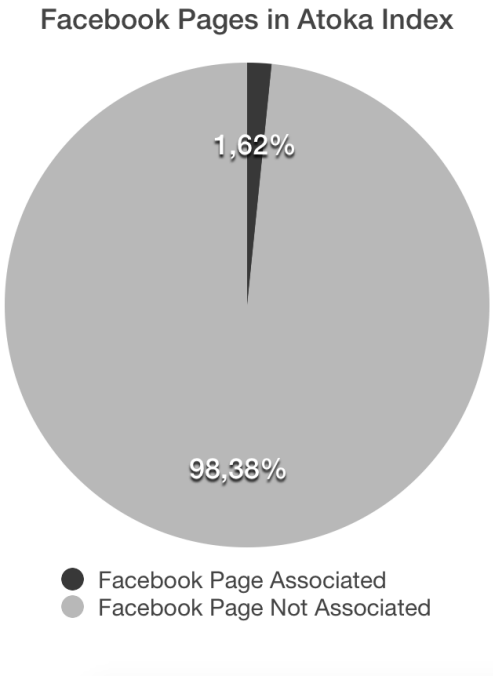

Figure 2.1: Only 96310 over 5952128 companies in Atoka had a Facebook Page or User associated before this project.

because in today's world it has become more and more important to have social information about companies if you want to find potential customers through a lead generation technique.

3.1.1 Data Sources

The data sources used in the project are mainly two: Atoka Index and Facebook's GraphAPI. Data is analyzed and collected in a single index, used by the classifier to predict if a Facebook page belongs or not to a particular company.

Atoka

Atoka is a collection of data coming from Cerved Group S.P.A and crawlers made at SpazioDati, stored in a Elasticsearch index. This elasticsearch index is the heart of atoka.io, the lead generation tool in which the results of this project will be shown, added to data previously available.

The most important fields of Atoka index used in the project are:

- ateco: ateco code of the company, representing the classification of economical activities
- description: a description of the company, used as a source for the research of keywords
- emails: list of email of the company
- entities: entities extracted with Dandelion API
- location: list of addresses of the company (companies may have more than a single office)
- social: social media link of the company (facebook, twitter, linkedin, etc.)
- website: official website of the company

It's important to underline the fact that not all the companies have a social and a website stored in Atoka index: the purpose, in fact, is to fill the missing values in social field.

GraphAPI

The Graph API is the primary way to get data in and out of Facebook's social graph . It's a low-level HTTP-based API that you can use to query data, post new stories, upload photos and a variety of other tasks that an app might need to do. The endpoint used for retrieving information is:

```
GET graph.facebook.com
  /search?
    q={your-query}&
    [type={object-type}](#searchtypes)
```

After being cleaned, the company name is put in the parameter q as q={your-query}, for each company pages and user are searched inside Facebook graph (because many companies still use user instead of pages). For making less number of GraphAPI calls and reducing the HTTP request time we used a batch request, collecting multiple calls in only one. The limit imposed by Facebook is a maximum of 50 call per batch request. In this project, every batch request was made with 50 request.

The Batch endpoint is:

```
curl \
    -F 'access_token={TOKEN}' \
    -F 'batch=[{ "method":"POST", \
                 "relative_url":"search", \
                 "body": {query_1} }, \
               { "method":"POST", \
                 "relative_url":"search", \
                 "body": {query_2}]' \
    https://graph.facebook.com
```

3.1.2 Problem Definition

Achieving the required functionality seems to be a three-step process.

First is to take companies' names from Atoka index, clean them and create a new elasticsearch index with Facebook responses. Facebook responses should contain the same information types of Atoka in order to compare them and predict which page should be the original one.

After having stored Facebook's information, the second step is to query the new index and find which local Facebook page may match with the single company. This step in very important because it shifts the problem from a "one-vs-all" classification to a "one-vs-few" classification. The query selects a small subset that may match with the single company and makes possible to compare the company with only 20-30 pages/users instead of all the dataset. Selected the company's pages/users, a metrics script has to define some feature of the pages, giving a list of n-uples representing how similar the page/user data is to the company one.

The last step is to train a classifier with the features computed in the previous step and then classify all the pages. At the end of the classification each company with a Facebook page should have the page linked by the classifier, with a percentage of how similar it is to that company's data. The classifier algorithm chosen is random forest. The choice was made for time reasons, for avoiding overfitting problems and because random forest is easier to be configured in an affordable way and amount of time.

One big problem is that some company name are not equal or contained in their Facebook page because the society name is only used for legal purpose and the real name is actually different. In that case it's impossible to find the page with Facebook's GraphAPI and the result cannot be contained in the resulting set of the classifier.

Bibliography

[1] Converting your profile into a facebook page. https://www.facebook.com/help/175644189234902. ultimo accesso 09/09/2015.