Davide Lagano
Georgios Patrikis

Lab Group 24

REPORT

Laboratory Exercise 2: Web Usage Mining "Light"

1. Introduction:

The second Laboratory Exercise aims to discover potentially suspicious books in order to find subversives. We will show how simple it is to collect information about users from the Internet without the need for strong programming skills and using tools that are freely available.
We will use Wishlists from Amazon and a freely available geocoding service. Moreover, We will work with a subset of approx. 6000 wishlists downloaded from Amazon.com. In this report, we will go through different steps: Keywords and matches; Sorting, linking and analysing; Finding addresses and mapping.

2. Discussion about tasks:

In the first task, we used the file “keywords.txt” containing a list of keywords. We will add the keyword “ISIS” in order to find books that could lead to subversives.
Accordingly, we research these keywords on the 6000 wishlists downloaded from Amazon.com, and we found the matches and the number of them. Surprisingly, we found 502 matches, that compared with the total number of the wishlists is quite a big percentage, equal to 8.4%. This number could make sense if the 6000 wishlists downloaded belong to suspicious people. But if they belong to random people, it could be considered only as a starting point, also because this number consider every wishlist in which at least one of the ‘suspected' keyword are inside. But obviously, these books can be informative.

After this preliminary phase, we want to deepen. Thus, we sort the matches that we found and through the “keywordlinks.pl” script we also create a directory for each keyword and links its files to this directory.
We can see that the word with the highest number of matches is ‘bible’, that count 324 matches. The word ‘isis’ count 130 matches, this means that for our purpose is a significant word, because compared with other words is a huge value.
An example of other number:
‘fahrenheit451’ count 6 match, ‘michaelmoore’= 8, ‘bravenewworld’ = 7, ‘torah’ =3. We also have to consider that not every book talks about the topic that we are considering. For example, in some book connected with the word ‘bible’, this word is used only to reinforce the concept of 'manual' or 'guide'.
As we mentioned before, “these books can be informative". This means that if we found just one book in the wishlist, we can't say that this person is suspected, because this book, for example, can explain the story of isis.Can be interesting to search for pattern or combination regarding the keywords.
Analysing the various HTML we noticed that only the first page was extracted from each wishlist, this is a problem as only the most recent books are analysed.
This greatly penalizes our research, as it is known that extremists are formed and study old books, following concepts that go back decades.

As a last step, we have the possibility of obtaining the city and residence status of the wishlist owner. We run the script ‘bookaddresses.pl’ and automatically we have a file .txt in each keyword directory with all the states and cities for the users who have given their shipping address for each keyword. We decided to analyse the file isis.txt. Exploiting the tool http://www.batchgeo.com/, we have mapped every address. The result of it is in the figure below.
We also decided to repeat the same procedure with the file ‘bible.txt’.
Comparing the two different maps, we noticed that the distribution of the addresses on the American territory are positioned in the same way. We can, therefore, conclude that we have not noticed anomalies because even though it may seem that the addresses are more distributed to the east, but after having reasoned it could be concluded that the distribution is proportional to the density of the USA population in the various states.

We can also conclude that it is possible to extract more detailed information about the users, like sex, age, date of birth, marital status. Afterwards, it could be useful to analyse them to discover an interesting pattern.

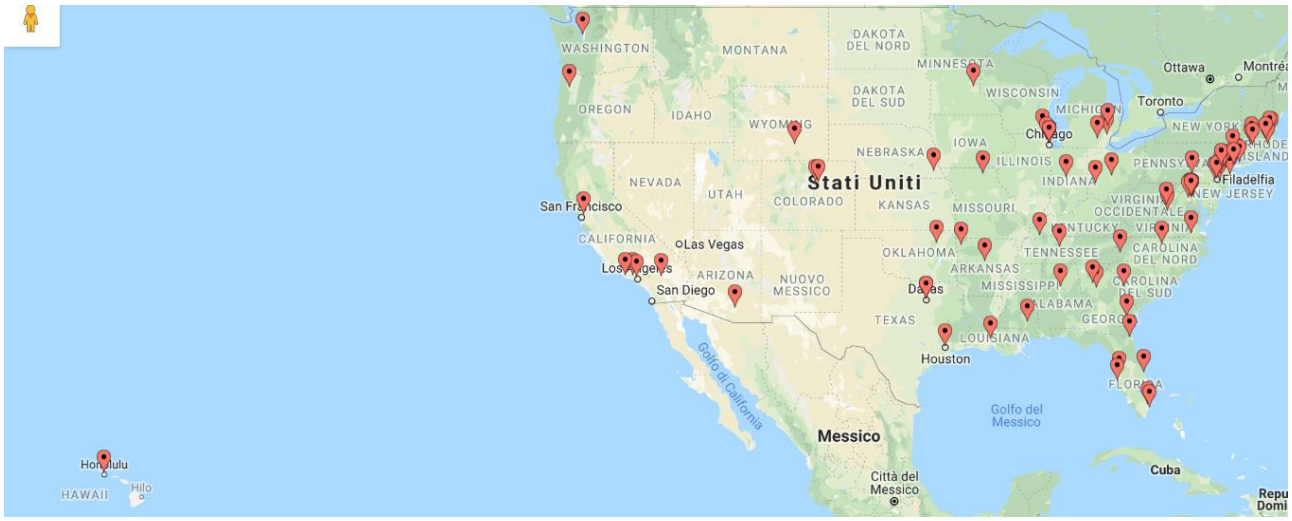

Figure 1

3. Conclusion

In this laboratory exercise, we learned how to obtain information, analyse and mapping them in order to get useful information.