Delft University of Technology

Bachelor Graduation Project

Initial Research Report

UrbanSearch

Authors:
Tom Brunnik
Marko Malis
Gijs Reichert
Piet van Agtmaal

Supervisor:
Claudia Hauff
Clients:
Evert Meijers
Antoine Peris

April 25, 2017

TU Delft

!!!!!!!!!!!!!!

Aspects: Algorithms Framework Component technology References [1]

Abstract

It is yet to be discovered how the importance of cities in the global network can be elucidated. In this paper, we develop a methodology to be able to reveal an answer to this matter. ...

Keywords: urban, city, data mining, document analysis, filtering

1 Introduction

Common belief is that agglomeration benefits are key to economic growth [1]. However it may be that this econmic growth's primary cause is the increase in (inter)national network embeddedness. We would like to further investigate this. Similar to research efforts in other domains such as financial trade [2] sales forecasting [3] and public health (Thornton et al., 2016) [**?**], the idea is to develop search queries that capture urban-urban interactions as they can be found on the web through the co-occurence of geographical names on websites e.g. "Zeeuws-Vlaanderen + Amsterdam " OR "Amsterdam + Zeeuws-Vlaanderen".

2 Requirements

2.1 Must haves

1. General

    (a) Adding city names
    (b) Grouping relations and "zooming" on these relations

2. Search Engine

    (a) Filter results
    (b) Data mining

3. Filtering

    (a) Logic Filters
    (b) Relations Filters

4. Machine Learning

    (a) Types of relations

5. Visualization

    (a) Statistics of relations? Query relations
    (b) Strength of relations
    (c) Types: ML CBS defined

2.2 Should haves

1. General

    (a) Pluggable datasets

2. Machine Learning

    (a) Generalising relations, grouping relations

2.3 Could haves

1. General

    (a) International city names

2. Visualization

    (a) Front end for the app

2.4 Would haves

3 Frameworks and Tools

3.1 Extraction

3.1.1 Information Sources

Common Crawl Common Crawl

Eurostat We identified Eurostat as a source that is not useful for the problem we're going to solve. Although Eurostat contains a lot of statistics on European cities, there is not enough useful information which contributes to giving more insight into the network connectivity of cities. Therefore, we did not include Eurostat as an information source.

3.1.2 methods

3.2 Filtering and Categorizing

3.2.1 Clustering

3.2.2 Filtering

3.2.3 Machine Learning

3.2.4 TF-IDF

basic idea: 1. using training data to assign values on words - filter meaningless words - assign words with highest value as categories? 2. Do the same on training data for each category (choose a few documents manually per category) and then check for websites for which categories has the highest value.

3.3 Search Queries

3.3.1 Enter Queries

3.3.2 Get Results

3.3.3 Specifications

3.4 Visualisation

3.4.1 neo4j?

3.4.2 Connection between cities

3.4.3 The Strength of these connections

References

[1] Michael E Porter. Location, competition, and economic development: Local clusters in a global economy. *Economic development quarterly*, 14(1):15–34, 2000.

[2] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using google trends. 2013.

[3] Lynn Wu and Erik Brynjolfsson. The future of prediction: How google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy*, pages 89–118. University of Chicago Press, 2014.