THE Heinz JOURNAL

Country Clustering Based on Search-Query Pattern Correlation

By:
Edgar Alejandro Anzaldúa Moreno
Riaz Esmailzadeh

Volume 9, Issue 1

Policy. Research. Practice.

The Heinz Journal
The H. John Heinz III College
Carnegie Mellon University
4800 Forbes Avenue
Pittsburgh, PA 15213

Country Clustering Based on Search-Query Pattern Correlation

Edgar Alejandro Anzaldúa Moreno and Riaz Esmailzadeh, Carnegie Mellon University, Australia

Executive Summary

This paper introduces a technique to cluster regions or countries for marketing purposes. It uses information gathered through aggregated website search-query patterns, available from web logs or analytics software. Clustering done using this technique helps grouping countries that similarly respond to a particular marketing campaigns potentially reducing marketing costs. The method collects aggregated geographical and search-pattern information from keywords and the visit volume of each keyword. It uses abstract keyword logs in web analytics software to identify and understand target market segments more effectively. It is argued that actual search logs could be better indication of market behavior than language or cultural boundaries. By analyzing web logs with this technique, marketers will be able to plan marketing campaigns strategically based on the empirical data on website visitors. This data is derived from huge long-tailed keyword search sources, and ultimately helps understand and adapt better to customer's interests. The paper starts with an introduction on the underlying problem, the analysis method, and its four basic steps: preparing the data, creating a correlation matrix, identifying criteria and region clustering using the Louvain method of community structure over random networks. The results for a specific dataset are then presented and conclude with a list of future work steps. Results in this paper show that country-based keyword analysis can yield search-query patterns that unveil connections between culturally dissimilar regions.

Introduction

Thanks to the internet, people all over the world, access thousands of webpages every minute. There are several ways of accessing a website, ranging from direct access (like typing an URL on the internet browser's address bar) to using internet search engines that facilitate access to approximately 18.63 billion web pages[1] through manual and crawler-based indexing techniques. Technology, standards and software tools together allow tracing of this kind of user behavior by enabling website owners to record how people get into their website: whether s/he types the URL into the address bar or if s/he used a search engine such as Yahoo, Google, Bing or Ask to get to the website. If a search engine is used, the actual search term is of interest: i.e. what was the keyword[2] that brought that user to the website.

Most internet users are likely to use a search engine when trying to find a website for the first time. Even when people know the URL of a website, they sometimes write the name of the site, the products or the services that the website offers in the search field of a given search engine to access them. Organizing this information in order to see how people perceive the company's

[1] Netcraft. Web Server Survey. August 2011. Netcraft | Internet Research, Anti-Phishing and PCI Security Services. 7 August 2011 <http://news.netcraft.com/archives/category/web-server-survey/>.
[2] Keyword for the purpose of this paper should be understood as search query, which may contain one or more keywords.

brand, product or services depending on their geographical location might enable companies in the future to further understand their customers based on their cultural background, nationality or the languages they speak. Moreover, marketing campaigns by companies with a wide variety of products or services usually group countries and/or regions based on a geographic or proximity approaches that might be biased by language, religion, or cultural perceptions.

Usual Hurdles Interpreting Large Dataset

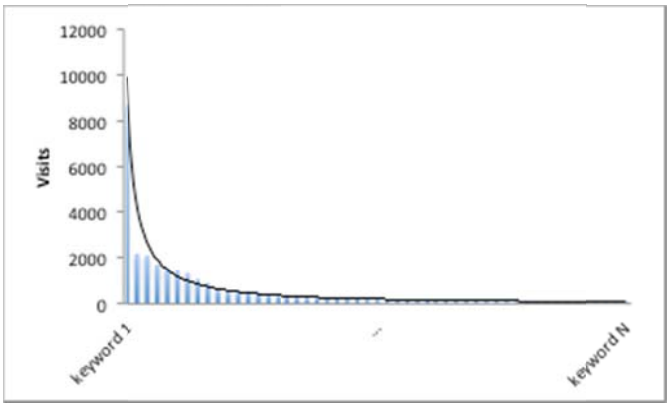


Figure 1. Example of how keywords logged by web analytics software could be charted; the shape of the curve resembles the long tail popularized by Chris Anderson[3]

Imagine the website of a large and renowned licensed office, home and server software development and manufacturer firm, or one of the top 10 universities in the world. Now imagine the traffic their customers generate on their websites through search engines; analyzing their customers' specific needs, could range from '*bulk licensing*' to '*instant messenger download*' or from '*postgraduate degrees*' to '*library resources*'. In each of these cases, the websites will log keywords: some with thousands of hits to some with one hit. Now imagine a large long-tailed keyword chart like the one shown in Figure 1 that compares keywords vs. visits, organized with those keywords that receive more visits to the left, and those that barely produce a visit to the website to the right.

The web log system might be able to break down the information further, making it possible to identify keyword visit volume per country. Figure 2 below depicts four different regions or countries using different keywords to access a website.

Column1	Region A	Region B	Region C	Region D	All Regions
keyword 1	11	201	52	60	324
keyword 2	66	22	54	123	265
keyword 3	20	96	56	61	233
⋮	⋮	⋮	⋮	⋮	⋮
keyword N	0	5	0	2	7

Figure 2. Comparison between four different regions that share the same search query patterns

[3] Anderson, Chris. The long tail: why the future of business is selling less of more. New York: Hyperion, 2006.

As shown in Figure 2 above, some regions may share the same keywords, but that does not mean that the keyword weighing[4] would be holding the same long-tail order for all of them. For example, *Keyword 1* may be the one that drives more traffic to the website across all regions in aggregate, but in individual regions, it is only true for Region B. As for Region A, the search query that gets the most traffic is *keyword 2*. This could be further visualized in Figure 3.

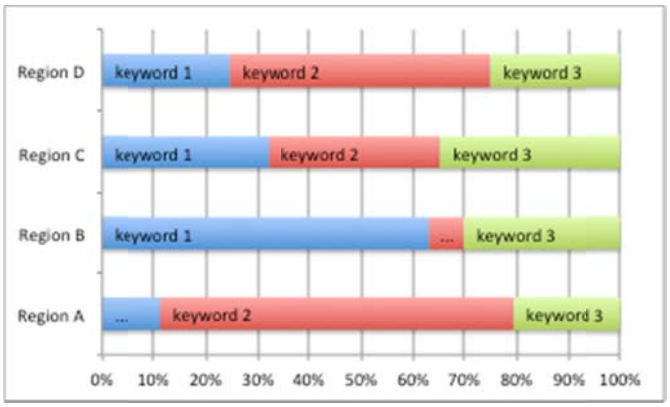


Figure 3. Show keyword weighting for different regions.

Figure 3 helps to visualize what search query is important for each region. For example, it could be assumed as per *keyword 2*, that Region A and D have the same interests to some extent as they share proportional similarities given that in both cases *keyword 2* is larger than *keyword 1* and *keyword 3*. The problem with Figure 3 is that it is unlikely that a company will deal with just three different keywords and four different regions. The datasets are expected to be comprised of thousands of different keywords and hundreds of different regions. As information becomes more granular, at the level of country or city, it becomes more difficult to order, understand, compare and ultimately extract information. Further, with time, the data aggregated by the web log becomes sparser. A full comparison between Region A and D is shown in Figure 4.

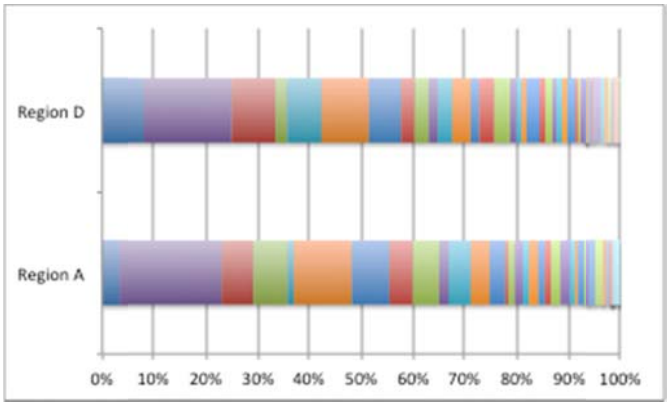


Figure 4. The chart above shows search-query pattern similarities and differences between two different regions (could be continents, countries or cities). Towards the rightmost side of the chart significant differences between both bars can be observed.

[4] Refers to the number of visits a website gets with one single keyword in a given period of time.

As is seen, the charts become more complex and difficult to compare as search query-patterns and regions appear. If we want to group all these regions based on how similar they are in terms of the properties discussed before, ordering and grouping these charts would become a time consuming task if based only on the basic charting and visual techniques used above, which are the most common and accessible analysis broadly known and found in commercial and open source analytics software.

To help visualize the information and grouping regions together according to their website visitors' interests, the next section proposes a clustering method to arrange and understand this data. Clustering is done in a way that would aid decision-making while segmenting target markets that have similar interests or behaviors captured in the search query patterns by web logs. In Section III an approach is described to accomplish this task.

Proposed Approach

We identify four big steps, as shown in Figure 5, to go through the whole process of clustering. For the purpose of this example regions are considered to be countries.

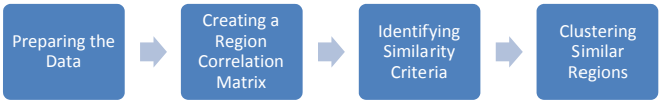


Figure 5. Proposed workflow to get the region clustering done.

Preparing the Data

Free web log software such as Google Analytics provide files in comma separated values (csv) format that can be manipulated easily with commercial and open source spreadsheet software. Figure 6 shows a typical file structure.

Keyword	Region	Visits
keyword 1	India	8703
keyword 1	Pakistan	8451
keyword 2	India	2216
keyword 1	Indonesia	2139
keyword 3	India	2139
keyword 1	Egypt	1818
⋮	⋮	⋮
keyword n	Mexico	2

Figure 6. Sample dataset format required to go through the clustering process. Column one (left) shows search queries that could be comprised of one or more words. Column two (center) holds the name of the region that a website got visited from, in this case countries. Column three (right) has the total number of visits that the unique combination between keyword and region sums.

To get rid of the potential noise[5] in the data set, non-relevant data to our analysis needs to be discarded. To do this a summary table needs to be created from the dataset that allows us to visualize rows as unique keywords and columns as different regions. The values held by this table would be the total number of visits that a combination between keyword and region has (See Figure 7); if a keyword does not exist for that region then it can be inferred the number is zero.

	India	Pakistan	Indonesia	Egypt	...	Mexico
keyword 1	1263	1047	491	1443	...	407
keyword 2	768	869	551	727	...	862
keyword 3	721	1486	418	490	...	633
keyword 4	130	571	978	0	...	0
keyword 5	353	1376	702	157	...	221
keyword 6	106	327	454	0	...	312
⋮	⋮	⋮	⋮	⋮		⋮
keyword n	0	1	1	1	...	0

0 Keyword does not exist for this region

Figure 7. This is how the summarized table should look like. Notice that every search query pattern (*keyword 1... keyword n*) should be different from each other. If the keyword does not exist for that region/keyword combination then it is filled with zero (as *keyword 4* in Mexico among others in this example).

Pearson Correlation

The following step involves creating a Pearson product-moment correlation coefficient matrix based on the values returned from the comparison between every column to all others from Figure 7.The Pearson product-moment correlation (Pearson Correlation) coefficient is used broadly by recommendation systems to measure the level of linear dependence between two different variables[6] Pearson Correlation's is given by the following formula:

$$corr(r_i, r_j) = \frac{\sum_{k=0}^{n}(v_{r_i,k} - \bar{v}_k)\left(v_{r_j,k} - \bar{v}_k\right)}{\sqrt{\sum_{k=0}^{n}(v_{r_i,k} - \bar{v}_k)^2}\sqrt{\sum_{k=0}^{n}\left(v_{r_j,k} - \bar{v}_k\right)^2}}$$

where r_i and r_j are the specific regions being compared, $v_{r_i,k}$ are the number of visits generated by keyword k in r_i, and $\bar{v}_k$ is the average number of visits generated by both regions r_i and r_j [6,7]. As the formula suggests, the correlation will only happen with those keywords that exist for both items (Figure 8, the row with the *tick* in *quality* column).

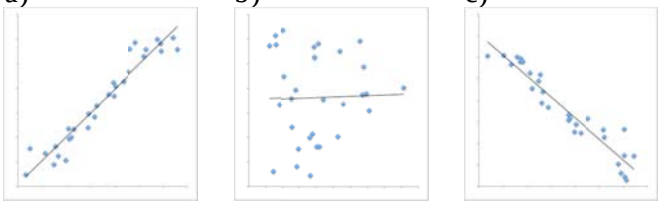
[5] By noise is meant the data that does not aggregates useful information to solve this problem.
[6] Melville , Prem and Vikas Sindhwani. Recommender Systems. IBM T.J. Watson Research Center. Yorktown Heights. <http://people.cs.uchicago.edu/~vikass/recommender.pdf>.
[7] Wikipedia contributors. Random graph. 12 July 2011. 7 August 2011 <http://en.wikipedia.org/w/index.php?title=Random_graph&oldid=439103301>

	Region A	Region B	Quality
keyword 1	41	11	✔
keyword 2	0	40	✖
keyword 3	52	0	✖
keyword 4	0	0	✖

Figure 8. The bigger the datasets are, and the greater number of rows containing non-zero regions being compared, the more significant and insightful information Pearson Correlation returns.

Pearson Correlation formula returns values in the range of $-1 \leq corr(r_i, r_j) \leq 1$; *one* (1) meaning that there is a complete correlation between the two regions, *zero* (0) where the datasets are uncorrelated, and *minus-one* (-1) where datasets are inversely correlated. Any number in between describes to what extent one dataset (r_i) is correlated with the other (r_j). Due to the nature of this exercise, negative correlation does not provide any meaningful information, as we cannot assume (as correlation does not mean causation) that increase in the use of one keyword might be due to the fact that a decrease in the use of another keyword is happening. This is because two countries' keywords and its visit volumes are assumed to be independent from each other. Figure 9 shows how correlated (a), uncorrelated (b) and inversely-correlated (c) datasets are displayed in a chart. Although it is unlikely that highly inversely-correlated datasets may appear, nevertheless, the nature of the equation might return negative values. As described above these are irrelevant in our analysis here, and therefore indicate that the correlation between those datasets is null.

a) b) c)



$corr(r_i, r_j) \approx 1$ $corr(r_i, r_j) \approx 0$ $corr(r_i, r_j) \approx -1$

Figure 9. Different correlation levels; *a* being a strong positive correlation near to one, *b* being almost a null or zero correlation and *c* a negative correlation of almost minus-one. It cannot be expected *c* to happen in search-query pattern analysis.

Before calculating the Pearson Correlation between all search-query pattern datasets, it should be clear that the dataset consists of many heterogeneous records to analyze. Due to the fact that most of our dataset has the shape of a long tail, it will remain true to the Pareto principle: Around 80% of the visits to a site happen in just 20% of the pages within it. Now for those pages, 80% of the visits are the result of 20% of the keywords, and so on. Given the large size of the dataset needed to be able to analyze this kind of information and its changes on the regular basis, it makes that Pareto's 20% important and with a noisy nature (as it will contain both valuable and non-valuable information that will be difficult to clean accurately). One way to clean the dataset, is to reduce unnecessary records. This can be done by removing those keywords which have resulted in only one visit for all the regions.

The next step involves spotting all those records that offer only one combination keyword/country, as they will not produce any significant information as per Pearson product-moment correlation coefficient. Weighing significance can be further analyzed if datasets are too small or if the results do not produce any insightful outcome (see Figure 8).

Creating a Regional Correlation Matrix

A correlation matrix is needed to arrange the Pearson Correlation comparison values resulting from comparing the search-query patterns from two different regions.

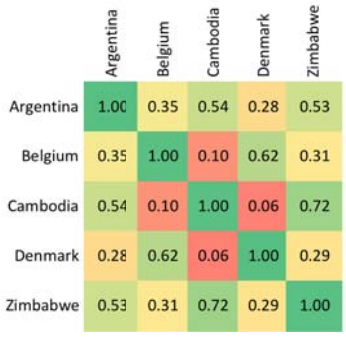


Figure 10. A correlation matrix between country-delimited regions ordered from A to Z in both axes based on keyword-driven visits to a website. Color has been added to help spot strong and weak correlation between regions.

A correlation matrix, in this case, will be a symmetrical bi-dimensional array of *n* regions in no particular order. Correlation between one country and itself should always equal 1 (see figure 10). When countries are sorted in the same order in both rows and columns as in Figure 10, diagonal symmetry is found.

One big problem that this specific correlation matrix has, is that values are not ordered in any way that could substantially aid region clustering, as values are scattered all across the matrix. Furthermore, there is no easy way to find the right combination of countries that maximize similarity and minimize the number of clusters.

Identifying Similarity Criteria

Once the Pearson Correlation matrix has been successfully created, identifying what is similar and what is not is needed. When arranging region columns from the most to the least similar (ordering top to bottom from largest to smallest the Pearson Correlation coefficient) we will start identifying to what extent a region is similar to what other regions. This is shown in Figure 11.

	India
India	1.00
Nepal	0.90
Sri Lanka	0.90
United Arab	0.88
⋮	⋮
El Salvador	0.05
Seychelles	0.02
Belarus	0.01
Tuvalu	0.00

	Philippines
Philippines	1.00
Malaysia	0.88
Sri Lanka	0.85
Indonesia	0.82
⋮	⋮
Senegal	0.04
Belarus	0.03
Tuvalu	0.03
Benin	0.00

Figure 11. Sorting correlation from largest to smallest from the correlation matrix for two different regions, India (left) and Philippines (right).

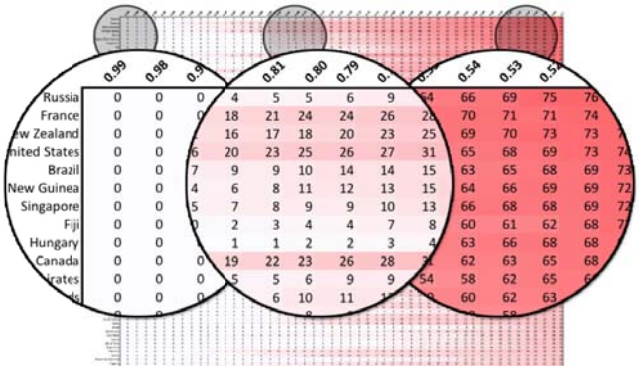


Figure 12. The table shows the number regions similar to the one in the vertical axis, for the Pearson Correlation Coefficient equal to or greater than the number showing on the horizontal axis. In this example, Russia is correlated with 5 other countries and France with 24, when $0.8 \leq corr(R) < 1$.

From the tables in Figure 11, it can be observed that there are few countries[8] that will correlate better with a specific country as compared to others. In the example above, India (left), correlates strongly with Nepal and Sri Lanka, and Philippines (right) with Malaysia and Sri Lanka. On the other hand, India correlates poorly with Belarus and Tuvalu; and Philippines with Tuvalu and Benin. Although both countries share some similarities, like Sri Lanka, Belarus and Tuvalu, finding or sorting all similarities between all countries can be a complex task. However, this provides us a hint on how to start grouping regions by similarity. This is shown in Figure 12.

We can infer from the table shown in Figure 12 that as the Pearson Correlation coefficient declines (reduce the comparison quality), the more countries have similarities with others, as well as if the Pearson Correlation Coefficient increases, it will reduce the number of countries that can be considered to be similar.

In this experiment, three different Pearson Correlation Coefficients were used, 0.9, 0.8 and 0.7. Results for value 0.9 returned just a handful of countries similar to others as opposed to 0.7 that returned too many. Hence, the coefficient picked (arbitrarily) was 0.8[9], The coefficient comparison suggests that anything bigger than 0.8 is considered similar, and anything below this number is not.

After the Pearson Correlation Coefficient that will return what is considered similar has been selected, creating the least groups of countries that are similar to each other is the next step.

.

[8] The paper has been talking about "regions" as the intended geographical granularity identifiable in our data set, but for the propose of the example regions is going to be narrowed specifically to "countries" instead.

[9] This number might vary from one case to another depending on how random[7] the network is.

Clustering Similar Regions

Based on the information collected so far and using graph theory[10] a region network can be modeled by linking together all those regions that have a Pearson Correlation Coefficient greater than 0.8 to identify possible region clusters.

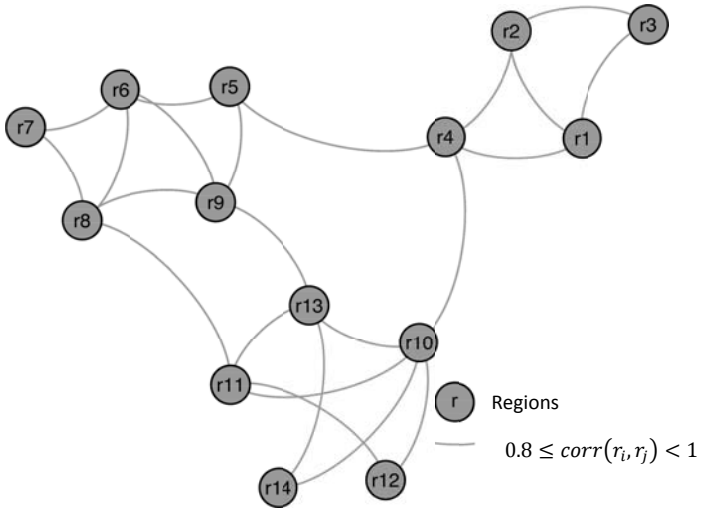


Figure 13. The figure shows an undirected graph[11] based on sample search-query patterns between regions (nodes) linked by its Pearson Correlation Coefficient (edges). In this figure region *r13* is linked by a coefficient larger than 0.8 with *r9*, *r10*, *r11* and *r14*.

The regions graph shown in Figure 13 can be broken into different partitions given the number of edges or links between the nodes or regions. This exercise uses the Louvain Method[12] to unveil *communities* or groups of regions that share similarities with each other on the search-query pattern-based country network; The Louvain Method uses Modularity Optimization to achieve this goal[13,14].

[10] Wikipedia contributors. Graph theory. 5 August 2011. 7 August 2011 <http://en.wikipedia.org/w/index.php?title=Graph_theory&oldid=443196082>.
[11] A graph in which edges have no orientation (Wikipedia contributors. Graph (mathematics). 7 August 2011 http://en.wikipedia.org/wiki/Graph_(mathematics)#Undirected_graph)
[12] Bastian, M. View Topic - What Is the Method of Finding Community Structure? August 7 2011 <http://forum.gephi.org/viewtopic.php?t=354>.
[13] Lambiotte, Renaud. Find Communities. 7 August 2011 <http://sites.google.com/site/findcommunities/>.
[14] Blondel, Vincent D, et al. Fast Unfolding of Communities in Large Networks. Universit´e catholique de Louvain. France, n.d.

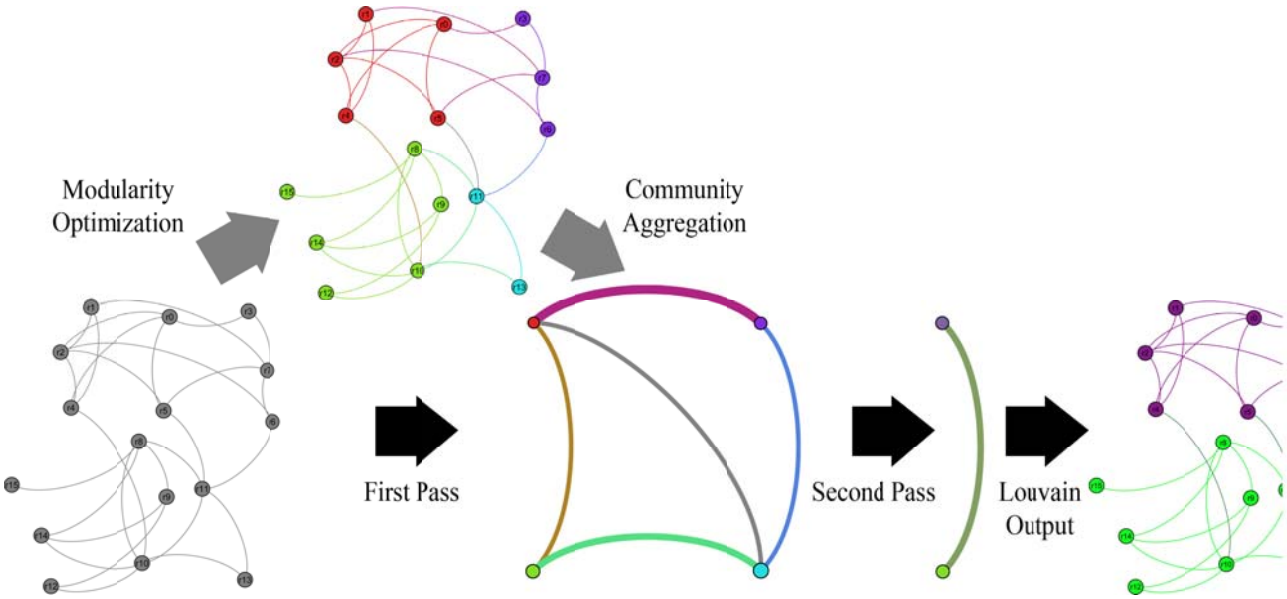


Figure 14. The figure explains how the graph gets broken down first into four different communities between neighboring nodes. Then those communities are taken as new nodes and two other communities are being built from them, until no more iterations can be achieved. The final output for the example shown in this figure has two communities. Adapted from the paper "Fast unfolding of communities in large networks" [10] by Vincent D. Blondel et al.

Briefly, modularity is defined by "the characteristic of a system that has been divided into smaller subsystems which interact with each other"[15]. Modularity maximization uses simulated annealing, spectral optimization and other algorithms that will break down a network structure in a graph into smaller communities. The use of heuristic algorithms is needed due to the intractable time consumed in evaluating all possible communities and comparing them with others. On the other hand the Louvain Method uses modularity maximization in small local communities that get aggregated and compared again as new nodes. This will be done until maximum modularity is achieved (Figure 14).

Remember that for the purpose of this paper, regions were assumed to be countries (represented by nodes in graph theory) and the Louvain method was successfully used to determine which countries are similar in terms of search query patterns. The next section presents the results achieved with real charts.

Results

As mentioned in section III of this paper, our dataset (like the one shown in Figure 6) consisted (after being cleaned) of 6005 records gathered over the course of three years. Those records where summarized in a table (like the one shown in Figure 7) consisting of 2029 different keywords for 123 different countries. Some of those countries were ruled out before and after the correlation matrix, due to noise cleansing and a poor correlation with all other countries. Most of the dataset used was in English for all countries. Cultural and geographical similarities were found in clusters made by this method but it was definitely not the rule. Some interesting links

[15] Modularity Definition. 2011 August 7 <http://www.pcmag.com/encyclopedia_term/0,2542,t=modularity&i=47187,00.asp>.

between countries were found regardless of those Cultural and Geographical proximity as shown in Figure 15. In the example shown in Figure 15[16], many expected country clusters can be found, like a cluster of Argentina, Chile, Colombia, Peru and Venezuela together. It can be concluded that this phenomenon originates due to the fact that they are all Spanish speaking countries using Spanish search patterns. But the fact that Turkey and Romania are together in the same cluster with them suggests otherwise. Furthermore, Spain and Mexico are together in another cluster with Brazil, USA and Canada and other countries from geographically distant regions. While comparing the clusters to the 2010 list of countries by Gross Domestic Product, we found that Mexico, Spain, Brazil, USA and Canada are grouped together too with a GDP larger than 1,000 billion US dollars. Argentina, Colombia, Venezuela and Turkey are together within the US$200-999 billion GDP group, and Chile, Peru and Romania in the US$10-199 billion GDP group. However, at this point there is no substantial evidence to prove a close relationship between GDP and product or service preferences based on search query patterns. But this could be one reason these countries are clustered together.

Figure 15. The graph for this figure shows a real example on how different countries could get clustered by Louvain Modularity and search-query patterns.

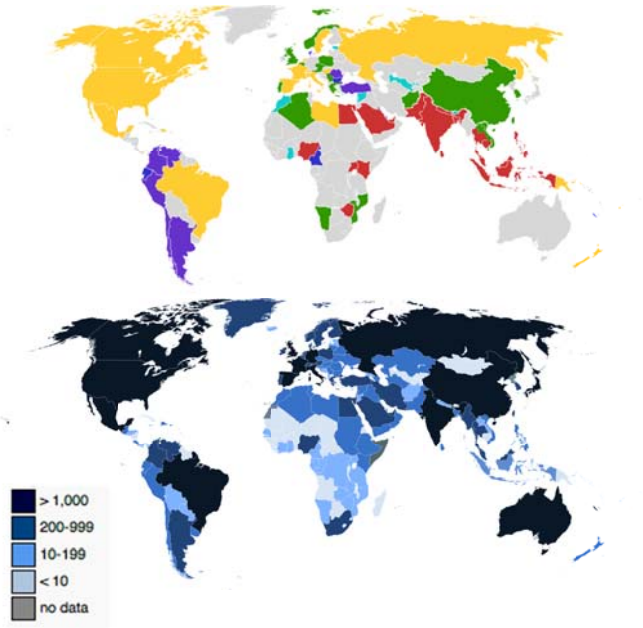


Figure 16. World maps showing country clusters from Figure 15 over the world map (top) and "nominal GDP of countries for the year 2010 according to the IMF [(bottom)]. Legend: (in billions of U.S. dollars)"[17]

After the completion of this method a couple of other approaches were set into motion. Both went through the same process, but one used webpage URLs instead of search query items, and the other used product sales. Both approaches were discarded since they could not return any significant correlation as there were not many clusters based on these particular similarities.[18]

[16] Australia was outside the scope of the study for this matter.
[17] Wikipedia contributors. List of countries by GDP (nominal). July 30 2011. August 7 2011 http://en.wikipedia.org/w/index.php?title=List_of_countries_by_GDP_(nominal)&oldid=442219977.
[18] This is given that the result graph consisted of one main group that contained almost all countries, and just a few other extremely small groups.

Future Work and Discussion

It should be noted that the whole correlation process does not consider the content of search-query patterns. Additional research should be undertaken in order to further analyse information from each country. For example, India, Philippines, Bangladesh and Malaysia have a lot in common, but we have not seen what information lies under that similarity. In this study, only Pearson Correlation Coefficients arrays and resulting similarities have been analyzed. The dataset contents, which have resulted in these coefficients, has not been considered. These represent some of the key items yet to be reviewed in order to enhance the outcomes of this study in the future. This paper only helps answer some questions on how to better understand the similarity between different countries, and helps answer questions about common interests that those countries might share. However, it is still a long way before a country's search query pattern lists can by synthesized and used to identify or predict user behavior in response to marketing campaigns. Recognizing similarities based on Pearson Correlation Coefficients is an important first step. It is useful to those entities with a large product or service portfolio, who would like to cluster regions based on search patterns garnered from web logs. With the rise of the semantic web, some of the hurdles mentioned in section V and VI can be overtaken as more insight into the actual search queries can also be found, giving rise to a new breed of web log aggregation systems.

Creating keyword categories could actually help understand the aggregated web log data. For example, it would enable marketers to identify how often a brand is used during a search query to get to the site, find how well positioned a product is in the market, etc.[19]. The use of the Term Frequency—Inverse Document Frequency (tf-idf) method on the aggregated search-query logs could also be used[20] to create keyword categories.

Conclusions

A method to group countries based on what people search on the Internet was proposed. This method enables a marketer to make an informed decision based on empirical data, rather than clustering based on non-empirical factors such as cultural links.

There are a few choices for software that provides insightful information to analyze aggregated web data. A popular tool is Google Analytics, a free and '*one size fits all*' solution. Because of its generality, it lacks specific functionalities to obtain tailored information. Fortunately Google Analytics provides sockets to extract updated information and the ability to export this information to databases or spreadsheet software.

As shown in this paper, clustering of data can be carried out with relative simplicity. The drawback is that it is time-consuming because it requires many different expertise areas. That is the Information Technologies department for large organizations should work together with the marketing department and information visualization experts to facilitate information, and ultimately create value for the overall organization through more efficient marketing campaigns.

[19] Murata, Tsuyoshi and Kota Saito. Extracting Users' Interests from Web Log Data. Tokyo Institute of Technology. Tokyo, n.d.
[20] Tonella, Paolo, et al. Using Keyword Extraction for Web Site Clustering. Centro per la Ricerca Scientifica e Tecnologica. Trento, n.d.

Ultimately this exercise tells us that there is more than meets the eye when it comes to aggregating search-query logs. Clustering information can provide input to formulating marketing strategies, and this should be of importance to many organizations.

Sources

Anderson, Chris. The long tail: why the future of business is selling less of more. New York: Hyperion, 2006.

Bastian, M. View Topic - What Is the Method of Finding Community Structure? August 7 2011 <http://forum.gephi.org/viewtopic.php?t=354>.

Blondel, Vincent D, et al. Fast Unfolding of Communities in Large Networks. Universit´e catholique de Louvain. France, n.d.

Lambiotte, Renaud. Find Communities. 7 August 2011 <http://sites.google.com/site/findcommunities/>.

Melville , Prem and Vikas Sindhwani. Recommender Systems. IBM T.J. Watson Research Center. Yorktown Heights.

Modularity Definition. 2011 August 7 <http://www.pcmag.com/encyclopedia_term/0,2542,t=modularity&i=47187,00.asp>.

Murata, Tsuyoshi and Kota Saito. Extracting Users' Interests from Web Log Data. Tokyo Institute of Technology. Tokyo, n.d.

Netcraft. Web Server Survey. August 2011. Netcraft | Internet Research, Anti-Phishing and PCI Security Services. 7 August 2011 <http://news.netcraft.com/archives/category/web-server-survey/>.

Tonella, Paolo, et al. Using Keyword Extraction for Web Site Clustering. Centro per la Ricerca Scientifica e Tecnologica. Trento, n.d.

Wikipedia contributors. Graph (mathematics). 7 August 2011 <http://en.wikipedia.org/wiki/Graph_(mathematics)#Undirected_graph>.

—. Graph theory. 5 August 2011. 7 August 2011 <http://en.wikipedia.org/w/index.php?title=Graph_theory&oldid=443196082>.

—. List of countries by GDP (nominal). July 30 2011. August 7 2011 <http://en.wikipedia.org/w/index.php?title=List_of_countries_by_GDP_(nominal)&oldid=442219977>.

—. Random graph. 12 July 2011. 7 August 2011 <http://en.wikipedia.org/w/index.php?title=Random_graph