The Geographical Life of Search

Ricardo Baeza-Yates*, Christian Middleton†, Carlos Castillo*
*Yahoo! Research; Barcelona, Spain
†Universitat Pompeu Fabra; Barcelona, Spain

Abstract

This article describes a geographical study on the usage of a search engine, focusing on the traffic details at the level of countries and continents. The main objective is to understand from a geographic point of view, how the needs of the users are satisfied, taking into account the geographic location of the host in which the search originates, and the host that contains the Web page that was selected by the user in the answers. Our results confirm that the Web is a cultural mirror of society and shed light on the implicit social network behind search. These results are also useful as input for the design of distributed search engines.

1. Introduction

The goals of this paper are three-fold. First, understanding how search engines are used from a geographic perspective is interesting on its own. For example, just confirming that linguistic or developmental factors are more important than geographical factors when studying inter-country similarity, gives insight on how services on the Web should be designed. Second, the search process can be seen as an implicit social network (e.g. people is related to people that search similar things [1]) and the geographical user behavior gives information about this social network and how society is reflected on the Web. Third, the search traffic among countries is interesting for the development of distributed search engines. In particular, the fraction of queries and clicked result pages that are local, gives information on where to locate a node in a distributed search architecture. Moreover, finding similar countries from a search perspective, enables a better design of the hierarchical organization of such a distributed architecture.

We analyzed data extracted from a sample query log from different points of view. The main objective is to describe how users behave, based on their location and the clicked URL, and test a set of hypotheses using the data obtained. This analysis represents from where a user need comes and where it is resolved, so in particular, implies traffic of information and transactions. So the geographical life of search is related to the geographical life of information, which is part of the social life of information [2]. In fact, this implicit social network is related to the Internet social network at large.

Our study addresses the goals above from the perspective of a particular search engine, Yahoo!. Hence, the results concerning the first goal are biased to the coverage and traffic of such search engine. Nevertheless, these results are valid for our second goal of understanding the social network behind search. On the other hand, the results on the second goal are an important piece of what a given search engine needs to migrate from a centralized replicated architecture to a truly distributed one.

In this study, among other findings, we observe that:

- The `.com` domain attracts a large share of traffic from several countries.
- Some generic top-level domains (gTLDs) are mostly used in the US, while others are used internationally.
- Vanity TLDs, which are country-code top-level domains (ccTLDs) used as if they were gTLDs, can be characterized by the traffic they receive and generate.
- Different countries have different rates of local search traffic, in which the searcher and the clicked page are in the same country.
- Countries in a similar geographic latitude or with a similar human development index tend to have similar traffic destinations.

Our findings mostly confirm what we expected to find. That is not strange, as today the Web is a mirror of society. Hence, our results are a confirmation of the geographical and cultural mirror, while in [3] we had an indication of the economical mirror.

The next section describes recent related work on this topic. Section 3 introduces the experimental framework we use. Section 4 presents our results for Generic Top-Level Domains (gTLDs) and Section 5 our main findings for Country-Code Top-Level Domains (ccTLDs). Section 6 analyzes the internal search traffic at the country and continent level and Section 7 the traffic that crosses country and continent level. Finally, the last section presents some concluding remarks.

2. Related Work

Understanding the underlying relation between Web structure and geographical features is an interesting research problem that has been studied recently. With some exceptions, most of the previous research on geographical aspects of the Web focuses on the contents and link of pages. Instead we look for insights on the interest of the Web users rather than on the structural linkage between the Web contents.

Usage analysis. Several works have studied the relationship between the query terms and the geographic location of users. Jansen and Spink [4] made an extensive study on the characteristics of Web usage of users in United States and Europe. They observed different behaviors between the US users and the European users, particularly in the way of structuring the query terms.

A study based on the most frequent geographic query terms used in Web search engines is presented by Sanderson and Han [5]. They observed that geography related terms are among the most frequently repeated words.

Gan et al. [6] investigated queries that use geographical terms to obtain location-specific results. Their results showed that geographical queries (*geoqueries*) tend to have more terms and geographical granularity (country, state, city) is closely related to the terms used. They also analyzed how different types of geoqueries were related to certain top-level domains.

Another approach has been used to try to determine the location of the users based on the query terms submitted to a search engine. Backstrom et al. [7] defined a probabilistic model that permits to infer the geographical center of a given search query based on a Web query log. This permits to understand the scope of a given query and study its geographical variation along time. Their study is fine-grained in terms that it points to specific geographical locations, while we aggregate the search traffic on the country and continent level.

Previous work was also done by Wang et al. [8] to determine the "dominant location" of a given query. Based on the search results and query logs, they are able to associate a geographical position to location-specific queries.

Hyperlink analysis. To understand the main features of Web structure at a hyperlink level, several studies have been done over different samples of the Web. The analysis done by Broder et al. [9] of a Web crawl permitted to identify the macro elements of the Web structure, as well as characterizing the in- and out-degree distribution of the Web pages. Baeza-Yates et al. [10] made a characterization of national domains by comparing 12 Web studies, covering 24 countries. They observed that the distribution of link-based metrics and degrees was consistent among the different countries. Also, they compared the results with cultural, linguistic, and economical indicators.

Bharat et al. [11] present a study of the structural linkage between Web hosts, based on three datasets from 1999, 2000, and 2001. They observed that there is a high geographical correlation between the link structure of hosts, followed by linguistic factors. Another important observation is that all host have the majority of links to other hosts within the same domain.

Based on the hostgraph of different countries, Baeza-Yates and Castillo [3] studied the relationship between commercial activities among countries and the link structure between hosts. They were able to observe a correlation between imports and exports of the given countries and the number of links between the hosts of each country-code TLDs.

Content Analysis. Another approach is using the contents of the Web pages to determine the geographical structure of the Web. Silva et al. [12] combine the geographical information extracted from the Web pages, during the crawling phase, with a graph-like structure to find locations. They found a correlation between the geographical location of a Web page and the pages being linked by it.

3. Experimental Framework

In this work, *host* refers to the unique name assigned to a server connected to the Internet, according to the structure of the Domain Name System (DNS)[13]. This structure was designed as a hierarchy of names where the upper level consists of a set of *Top-level Domains* (TLD). The top-level domains can be separated into two main groups: *Country-code top-level domains* (ccTLD), and *Generic top-level domains* (gTLD). The ccTLDs are a set of two letter country codes associated to each country according to ISO 3166-1[14], while the gTLDs are a set of general-purpose domains such as `.edu`, `.com`, `.net`, `.org`, etc.

In this paper the statistical median of a variable x is represented by $\tilde{x}$, its standard deviation by σ, and $H(x)$ represents the entropy of variable x.

3.1. Traffic Graphs

To represent the traffic among countries and domains we use two types of graphs. Domain traffic graphs are bi-partite graphs indicating the fraction of all clicks from searchers located in a country to URLs located in domains. Country similarity graphs are undirected graphs reflecting the similarity between two countries in terms of their traffic destinations.

Country-domain traffic graphs. To represent the traffic observed by the search engine, we use a country-domain graph such as the one depicted in Figure 1. This graph $G = (V, E)$ has a set of nodes $V = C \cup C' \cup D$ where C is a set of countries, C' the corresponding set of ccTLDs for those countries, and D a set of gTLDs. There is a bijection $c : C \rightarrow C'$ from each country to its corresponding ccTLD. The graph is bipartite and the set of edges is $E \subseteq C \times (C' \cup D)$.

A matrix $\mathbf{W}_{|V| \times |V|}$ represents the number of clicks in the country-domain graph, where w_{ij} is the number of clicks by users in the country $i \in C$ on documents in the domain $j \in C' \cup D$. This traffic is *incoming* for the domain j, and *outgoing* for the country i. All the countries generating the traffic received by a domain j are called the *traffic sources* of j, all the domains that receive traffic from a country i are the *traffic destinations* of i. Furthermore, we name *intra-country* to all the traffic from country i to its corresponding domain $c(i)$, and *inter-country* to the traffic from a country i to a domain $c(j)$ $\forall j \neq i$.

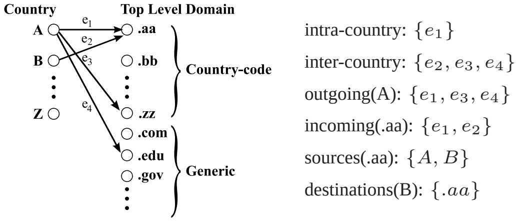


Fig. 1. Example of incoming and outgoing traffic in the bipartite graph.

By aggregating the countries in C into their corresponding continents, we can define a *continent-domain graph* that shows the traffic between continents and domains. The definitions used for the country-domain graph can be extended trivially to this graph.

Country similarity graphs Based on the traffic information, it is possible to define a similarity function between two countries (or continents) using the common domains clicked by the users, and create a *country-similarity* graph.

We can define a *country-similarity* function between the countries, based on the traffic information found in the matrix **W**. For each country i, we normalize their outgoing traffic (w_i), such that $\sum_k w_{i,k} = 1$. Finally, we define the country-similarity of two countries i and j as the cosine of their normalized outgoing traffic w_i and w_j .

This definition can be extended to create the *continent-similarity graph*, where each node corresponds to a continent, and the similarity is based on the aggregated traffic of the countries belonging to the continent.

3.2. Query Log Processing

Our base data is a large uniform sample of the Yahoo! search engine in early 2008. This is a log of all the actions of a set of users in the search engine during a certain period of time; essentially, the *queries* users submit and the *clicks* on URLs in the result sets. Our sample contains the query, user location (at country level), timestamp, and clicked URL of the request submitted by the user, among other attributes. Since we were interested in analyzing the domains clicked by the user, we filtered the URLs that were identified only by an IP address and had no corresponding domain name associated to them. The main reason for this filtering was because we are also looking for the relationship between the ccTLD of a URL and the location of the hosting server, hence the IP alone is insufficient information for our study.

As a result, we obtained a set of 840M clicked queries. Additionally, each clicked URL was parsed to extract the corresponding top-level-domain. To filter out noise from our observations, we eliminated the *inter-country* traffic that was below a certain threshold and corresponded to very few clicks. This threshold was obtained by analyzing the cumulative traffic from each country to other countries and discarding the last 0.01% of it.

3.3. Geolocating Hosts

From the 840M clicked URLs obtained from the query logs, we extracted a list of the most frequent unique hosts, and made a DNS-lookup on each of them to obtain the IP address of the server hosting the site. After discarding the hosts that could not be DNS-resolved, we obtained 759,153 unique hosts, where 593,433 hosts belonged to a gTLD and 165,720 hosts belonged to a ccTLD.

Next, using the IPligence [15] database, each IP address was mapped to the country were its server is located.

3.4. Country Information

We analyzed possible relations between the traffic among countries and their corresponding demographic information. For doing this, we extracted 24 features for each country from The CIA World Factbook and the UN Human Development Report 2007/2008. They correspond to statistical data such as population, area, life expectancy, etc; sources and a complete list of them is presented after the references. These attributes are used to calculate the correlation between them and the traffic similarity of countries.

For the mapping from countries to ccTLDs we followed ISO 3166-1 plus a few exceptions as sometimes the country and the domain does not match in an strict sense, but in practice they do match in their usage. One example is Great Britain where most people use the `.uk` domain and not the `.gb` domain.

To associate each country to a continent, we used the commonly adopted definition of 7 continents: Antarctica (AC), Africa (AF), Asia (AS), Europe (EU), North America (NA), Oceania (OC), and South America (SA)[1]. Notice that as the country domain of the main country in the Internet (US) is not used (.us), the US does not appear in many of the results. We plan to extend this study using the geolocation of the URL to include the US. Also we can precise better the origin of the search using the geolocation of the searcher, although the person can be a tourist and hence this assumption breaks down. So, for now we are assuming that the starting point of the search is a good proxy for the location of the searcher. In the tables that come later, we will refer to the continents using their abbreviation.

4. Generic TLDs

In this section we study the traffic and location of Generic TLDs (gTLD). This analysis can help to understand how people actually use these domains.

4.1. Traffic to the .com Domain

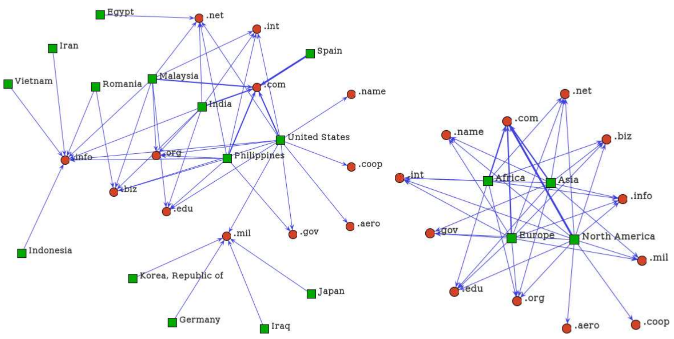
The `.com` domain stands out in our dataset as the most used domain for hosts and the one that receives the larger share of traffic, hence making it interesting to analyze separately.

Analyzing the traffic sources of `.com` we observed that there are 175 countries (of a total of 232) that have clicks to the `.com` domain.

We observe that most countries have at least 2/3 of their traffic to `.com` and even the countries where searchers click less on a `.com` domain, have more than 45% of their traffic to this domain.

Also, we can observe that, although most of the countries have the majority of their traffic to the `.com` domain, only a few of them are a relevant traffic source for this domain. We can observe that the `.com` domain is mainly influenced by countries in North America, Middle East, South East Asia, and part of Europe. Only 12 countries contribute individually more than 0.5% of the total incoming traffic to `.com`: United States, Philippines, Malaysia, India, Spain, Canada, Great Britain, Indonesia, United Arab Emirates, Egypt, Romania, and Iran. Many of the countries in this list have a significant percentage of `.com` hosts in their own country, such as Canada, Spain or the UK.

[1] We include Central America and the Caribe in North America. This would not be necessary in the European tradition of 6 continents where America is just one continent.



(a) Country-level (b) Continent-level

Fig. 2. Domains with significant traffic from each (a) country and (b) continent; only gTLDs.

4.2. Traffic to other gTLDs

Figure 2(a) (graph created using `JUNG`[2]) presents the traffic to gTLDs from their largest traffic source. We filtered the graph to include only countries that individually contributed at least 1% of the total incoming traffic to the gTLD. We can observe that the traffic to the largest gTLDs (i.e., `.com`, `.edu`, `.net`, `.org`, and `.biz`) is generated from United States, Malaysia, Philippines, Romania, India, Spain, and Egypt. Some gTLDs receive almost all their visits only from very few domains: `.coop`, `.name`, and `.aero` are only reached by searchers from United States; `.biz` and `.info` from Asian countries and Romania. A different distribution is observed in the `.mil` domain that is reached by searchers from the United States, Japan, South Korea, Irak, and Germany. This can be due to the location of US military bases in Asia and Europe.

We compared the traffic from each continent to the gTLDs, also considering only the traffic that represented at least 1% of the traffic destinations for each continent. This is represented in Figure 2(b). We can observe that for all the continents a large share of their clicks goes to the `.com`, `.edu`, `.org`, `.gov`, and `.net` domains. The `.aero` and `.coop` domains are mostly interesting to searchers from North America only.

4.3. Which Countries Host gTLDs?

Most of the hosts in the gTLDs are hosted in the United States. We analyzed the distribution of the countries in which the hosts corresponding to each gTLD are located. We show the entropy ($H(x)$) of this distribution in Table 1.

gTLD	$H(x)$	Top-1	Top-2	Top-3	Top-4
`.info`	2.21	0.61	0.11	0.08	0.06
`.net`	1.97	0.71	0.09	0.04	0.03
`.biz`	1.85	0.70	0.09	0.05	0.05
`.com`	1.57	0.77	0.07	0.04	0.03
`.org`	1.52	0.78	0.08	0.03	0.03
`.mil`	0.93	0.82	0.11	0.04	0.01
`.gov`	0.30	0.96	0.03	$<$0.01	$<$0.01
`.edu`	0.26	0.98	$<$0.01	$<$0.01	$<$0.01

TABLE 1. Entropy of the location of the countries hosting a gTLD and the percentage of their top-4 countries.

Most gTLD domains are concentrated in the United States, which is the Top-1 column of Table 1, but some are more highly concentrated than others. For instance, `.gov` and `.edu` have an entropy close to zero meaning that basically all of them are hosted in only one country. The hosts in `.biz`, `.net`, and `.mil` are more spread geographically; Figure 3 shows the cumulative distribution of countries hosting each of the gTLDs.

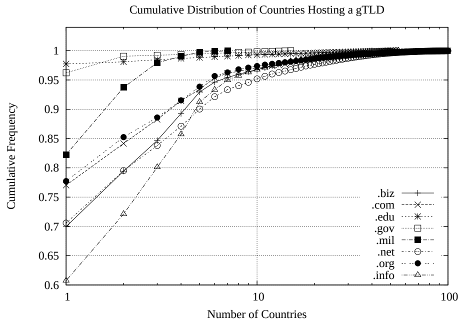


Fig. 3. Cumulative frequency of countries hosting gTLDs.

A possible explanation for this is that the first group of domains were designed to host governmental, educational, and military sites, and the registration entities have established restrictions on the sites that can apply for one of these domains (with few legacy exceptions). Since these domains generally need to be associated to their corresponding country, this division has been commonly moved to a secondary domain on a ccTLD (e.g., `.gov.uk`, `.gc.ca`, `gob.cl`). The rest of the gTLDs were designed to host sites that do not need to be associated necessarily to a particular country but to a broader area of use, hence the distribution into countries is less concentrated.

5. Location of Servers and Country TLDs

This section focuses on the analysis of the relationship between the geographical location (in the following geolocation or location) of the server hosting a domain and the country-code top-level domain (ccTLD) used to identify the server.

5.1. How can We Characterize "Vanity" TLDs?

According to [13], the main purpose of a ccTLD is to group together the domains of a country, while gTLDs should be used to group together domains based on a general category of organizations. However, some ccTLD are used as gTLDs due to commercial or phonetic characteristics of their code. For instance, the `.tv` (Tuvalu), `.fm` (Federated States of Micronesia), or `.am` (Armenia) domains are used by companies related to television and radio. These type of country-code TLDs are referred as a *Vanity ccTLD*.

Vanity ccTLD are clearly outliers with respect to several statistical properties. First, they easily stand out when looking at the *relation between outgoing and incoming traffic*.

Vanity score. We define the *vanity score*: $vanity(j)$ of a country-code domain j as:

$vanity(j) = 1 - \sum_{k \in C'} w_{c^{-1}(j),k} / \sum_{k \in C'} w_{c^{-1}(k),j}$, where $c^{-1}(j)$ recovers the country of the domain ccTLD j. We can observe

[2] http://jung.sourceforge.net/

Domain - Country (Cont.)		Score	Domain - Country (Cont.)		Prob.
`.tv` Tuvalu	OC	∼1.00	`.bz` Belize	NA	0.01
`.cc` Cocos Islands	OC	∼1.00	`.fm` Micronesia	OC	0.01
`.nu` Niue	OC	∼1.00	`.la` Laos	AS	0.04
`.ms` Montserrat	NA	∼1.00	`.li` Liechtenstein	EU	0.07
`.ws` Western Samoa	OC	∼1.00	`.ag` Antigua	NA	0.08
`.tk` Tokelau	OC	0.99	`.ug` Uganda	AF	0.09
`.tm` Turkmenistan	AS	0.99	`.ws` Western Samoa	OC	0.09
`.fm` Micronesia	OC	0.99	`.am` Armenia	EU	0.11
`.to` Tonga	OC	0.98	`.bd` Bangladesh	AS	0.11
`.sh` Saint Helena	AF	0.98	`.mu` Mauritius	AF	0.15
`.st` Sao Tome & Principe	AF	0.96	`.bi` Burundi	AF	0.17

(a) Vanity Score (b) Prob. ccTLD hosted same country

TABLE 2. (a) ccTLDs with Vanity score > 0.95. (b) ccTLDs in which less than 20% of the sites are hosted in the corresponding country.

that if $vanity(j)$ is high, it may indicate that the domain is being used as a vanity ccTLD, since the outgoing traffic is insignificant compared to its incoming traffic. We observed a clear separation at $vanity \geq 0.9$, which might indicate a behavioral difference of the domains. As shown in Table 2, these domains can have alternative uses due to its similarity to an abbreviation (`.tv`, `.fm`, `.ws`, etc.), its phonetics (`.nu`, `.to`, etc.), they offer free hosting in exchange of displaying advertisement on the hosted sites (e.g. `.tk`), or the domain belongs to a country with low population, hence having many names available (e.g. `.ms`).

For each country, we analyzed the probability that a ccTLD site is hosted in the same country as the one described by its ccTLD. We observed that the probability was non-uniform and tends to be close to one (median of 0.66), while there exists a group of ccTLD where very few of their hosted sites are located in their corresponding country. This may indicate that they are being used as vanity ccTLDs. Table 2(b) presents a list of the domains where less than 20% of the clicked sites are located in their corresponding country, in some cases because of vanity domains with an easy to identify abbreviation (e.g. `.am`, `.ws`, or `.fm`), and in other cases because of a lesser development of the Web in the country.

6. The Internal Search Traffic of a Country

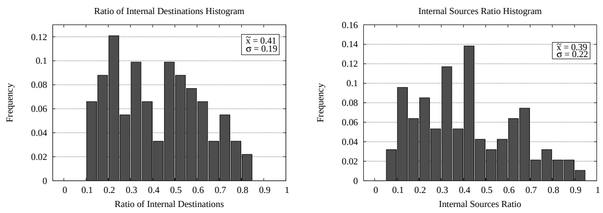
In this section we analyze the *internal search traffic*: visits to pages in which both the searcher and the clicked page are in the same country.

We divide this analysis into two parts: first we analyze the visitors from a country and their search traffic destinations, and then the hosts of a country and their search traffic sources.

6.1. Ratio of Traffic to Internal Destinations

We define the *ratio of internal destinations* (r_i) of a country as the probability that a user, after submitting a query, clicks on a site of its corresponding ccTLD, given that he/she has clicked on a ccTLD. The internal-destinations ratio of a country i is defined as: $r_i = w_{i,c(i)} / \sum_{k \in C'} w_{i,k}$.

Figure 4(a) presents an histogram of the distribution of the internal-destinations ratio for all the countries in the query log file. We can observe that most of the countries have a ratio of less than 1/2.



(a) Internal-destinations (b) Internal-sources

Fig. 4. Histograms of (a) internal destinations ratio of countries, and (b) the internal-sources ratio of domains.

There are few countries where more than 1/2 of their traffic (to a ccTLD) is directed to their corresponding country. They are shown in Table 3. Apart from countries belonging to the British Commonwealth (Great Britain, Australia, Canada, and New Zealand), the others have a high internal-destinations ratio possibly because of linguistic considerations.

Country (Cont.)		Ratio	Domain (Cont.)		Ratio
Brazil	SA	0.82	`.my` Malaysia	AS	0.94
Vietnam	AS	0.81	`.ro` Romania	EU	0.89
Poland	EU	0.79	`.us` United States	NA	0.89
Romania	EU	0.78	`.ir` Iran	AS	0.85
Great Britain	EU	0.77	`.vn` Vietnam	AS	0.81
Malaysia	AS	0.75	`.dz` Algeria	AF	0.80
Hungary	EU	0.73	`.id` Indonesia	AS	0.78
Australia	OC	0.72	`.ph` Philippines	AS	0.76
Russian Fed.	EU	0.72	`.pr` Puerto Rico	NA	0.72
Denmark	EU	0.71	`.ge` Georgia	EU	0.72

(a) Internal-destinations (b) Internal-sources

TABLE 3. (a) Countries with high internal-destinations ratio. (b) Domains with high internal-sources ratio.

We can also observe from this list that the geographical distribution of the countries with high internal-destinations ratio is not equally distributed. For this reason, we separated the analysis into continents, this is shown in Table 4. Three groups can be identified: countries in Oceania (basically Australia) have a high internal-destinations ratio ($\tilde{x} \approx 65\%$); Europe, South America, and Asia have a medium ratio ($\tilde{x} \approx 40\%$); Africa and North America have a low ratio ($\tilde{x} \approx 25\%$).

6.2. Ratio of Traffic from Internal Sources

Another important characteristic is the converse of intra-country clicks, from the point of view of sites: *ratio of internal-sources* of a country. This is defined as the probability that a visitor to a host in a ccTLD is located in the same ccTLD as the host.

A high ratio of internal-sources may indicate that the contents of the Web pages of that domain are of interest, mainly, to their nationals. The *internal-sources ratio* (q_j) of a domain j is defined as $q_j = w_{c^{-1}(j),j} / \sum_{k \in C} w_{k,j}$.

Figure 4(b) shows an histogram of the internal-sources ratio of the domains in the query log and Table 3 presents a list of the countries with the highest ratio (greater than 0.7). We can observe that `.my` (Malaysia), `.ro` (Romania), and `.us` (United States) domains are visited almost exclusively (90%) by people in their corresponding countries. Also, most of the countries that have hosts that depend heavily on the traffic

Continent	$\tilde{x}$	σ	Continent	$\tilde{x}$	σ
1. Asia	0.55	0.25	1. Oceania	0.67	0.23
2. Africa	0.50	0.21	2. Europe	0.51	0.17
3. North America	0.43	0.21	3. South America	0.42	0.19
4. Europe	0.33	0.20	4. Asia	0.37	0.21
5. South America	0.33	0.09	5. North America	0.27	0.15
6. Oceania	0.15	0.20	6. Africa	0.25	0.17

(a) Internal-referrals (b) Internal-destinations

TABLE 4. (a) Median values of internal-referrals ratio. (b) Internal-destinations ratio, sorted by continents.

from their nationals are located in Asia, most probably due to particular languages.

By separating the analysis into continents, as shown in Table 4, we can identify three groups. Hosts in Asia, Africa, and North America have a high ratio of internal-sources (median $\tilde{x} \approx 50\%$); Europe and South America have medium ratio ($\tilde{x} \approx 30\%$); and Oceania (Australia) has low ratio ($\tilde{x} \approx 15\%$). As observed in Table 3, Asia concentrates the largest number of domains whose main traffic originates from their corresponding country (languague issues again).

In general, there seems to be little correlation between the internal-destinations ratio for visitors of a country, and the internal-sources ratio for hosts in that country. In fact, the fraction of outgoing traffic that stays in the country, and the fraction of incoming traffic to sites that originates in the same country, seem to be mostly independent.

7. Traffic among Continents and Countries

This section analyzes the traffic among different countries. This can help to understand which countries have major influence on other countries, as well as understanding which countries share common interests.

7.1. The Traffic Among Countries

We first studied country-level domains with a large share of traffic from another country. Figure 5 presents those countries, among which the two most important ones are the `.uk` (United Kingdom) and `.ru` (Russian Federation). This can be explained mainly by historic and linguistic relationships between those countries and the countries that contribute to their traffic.

For this analysis, we only considered the domains that represented at least 4% of the external outgoing traffic of each country and for visualization purposes, we omitted 60 countries that only had intra-country links.

We also studied the entropy of the sources and destinations of the traffic for all countries. A country with high destinations-entropy indicates that people from a given country are interested in the contents of pages from multiple domains. A country with high sources-entropy indicates that the contents of the sites hosted in the corresponding domain are of interest to people in multiple countries. We considered only ccTLDs in this analysis. Table 5 shows the countries with highest sources- and destinations-entropy, respectively. We can observe that Asian and European countries have the most diversified traffic.

Figure 6(a) presents a scatter plot of the relation between the destinations- and sources- entropy. There is a weak correlation between these variables ($\rho = 0.241$).

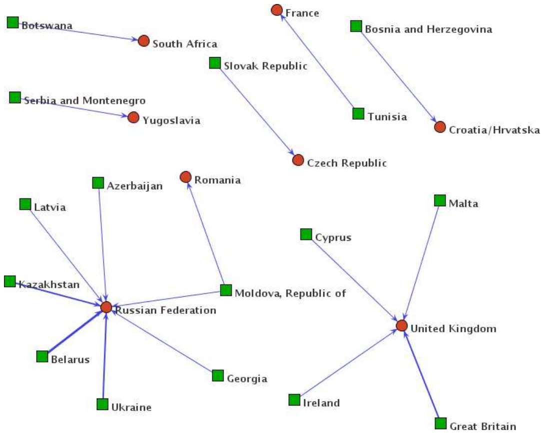


Fig. 5. Main groups of domains with significant traffic from each country (without gTLDs).

Country - Continent		Entropy	Country - Continent		Entropy
Iraq	AS	4.67	Moldova, Republic of	EU	2.87
East Timor	AS	4.23	Uzbekistan	AS	2.79
Grenada	NA	4.17	Vatican City	EU	2.77
Lebanon	AS	4.17	Kazakhstan	AS	2.74
Antarctica	AC	4.13	Belarus	EU	2.72
Nepal	AS	4.12	Ukraine	EU	2.71
Afghanistan	AS	4.09	Kyrgyzstan	AS	2.70
Zimbabwe	AF	4.08	Tajikistan	AS	2.63
Ivory Coast	AF	4.08	Azerbaijan	AS	2.63
Maldives	AS	4.07	Slovak Republic	EU	2.61

(a) Sources entropy (b) Destinations entropy

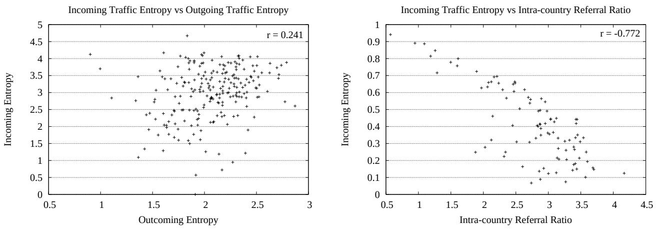
TABLE 5. List of top-10 countries (+Antarctica) with the most diverse traffic sources/destinations.

We studied the correlation with the ratio of internal- destinations and internal-sources. There exists an inverse correlation ($\rho = -0.772$) between the incoming traffic entropy and the internal-referrals ratio, as shown in Figure 6(b). Naturally, whenever a country-code domain has a narrow set of traffic sources, the country itself is one of the most important among those sources.

7.2. The Traffic among Continents

By aggregating the countries into their continents, it is possible to find relations at a broader level.

Figure 7 shows the most visited ccTLDs from each continent. We considered only the traffic that represented at least 1% of the traffic destinations for each continent. Interestingly, the resulting figure is a tree, even when it was drawn without further pruning.



(a) Entropy-destination (b) Entropy-traffic

Fig. 6. Scatter plot of (a) entropy of destinations and sources of traffic for each country, and (b) entropy of traffic-sources versus the ratio of internal-sources for each country.

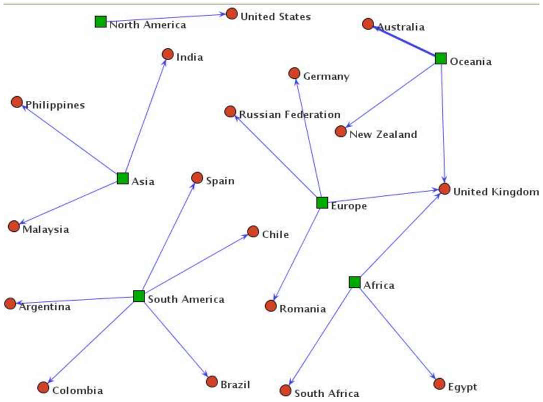


Fig. 7. Domains with significant traffic from each continent (without gTLDs).

We can observe that the `.uk` (United Kingdom) domain is frequently visited from Europe, Oceania, and Africa. South America has its traffic sources more diversified into several domains, while the rest of the continents have their sources concentrated in few domains. Another important observation is how Spanish sites (`.es`) have a greater influence in South America, where most of its former colonies are, than in Europe.

7.3. Traffic Similarity

Another type of usage analysis can be done by finding the similarity between two geographical regions based on their traffic destinations. This analysis can be made at different granularity levels (country, continent) to identify possible groups of regions that show similar traffic patterns.

Country similarity. Figure 8 presents a force-directed graph of the most similar countries, based on the similarity of their traffic destinations. First, we defined a vectorial space where each axis corresponds to a ccTLD, and a country can be represented in this space, based on their relative traffic to the ccTLDs. Next, we computed similarities between the countries as we explained in 3.1. The graph was filtered out to only show the edges where the similarity is larger than 0.95.

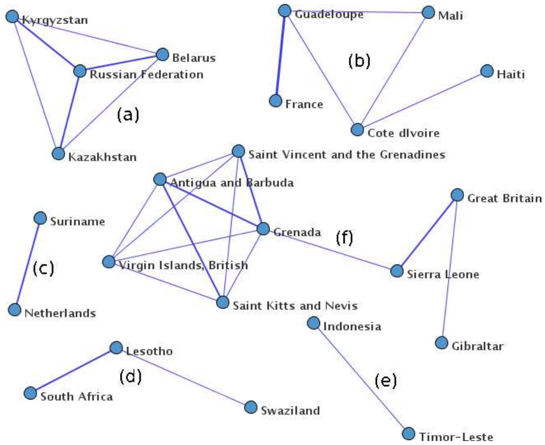


Fig. 8. Similar countries ($sim \geq 0.95$).

We can identify 6 different clusters of countries, where half of them have 4 or more countries. These clusters can be explained by a combination of linguistic and geographic characteristics: family of Indo-European language (Figure 8a), French speakers (Figure 8b), Dutch speakers (Figure 8c), Swazi speakers (Figure 8d), Tetum speakers (Figure 8e), and English speakers (Figures 8f).

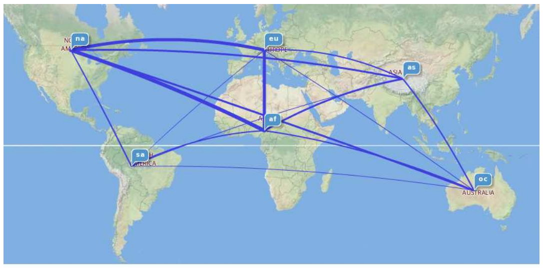


Fig. 9. Similarity between continents.

There are many possible features (characteristics) that can help explain the similarity between countries. Our analysis is based on the demographic features presented in Section 3.4. To determine which of these features could describe better the relation between countries, we used *Laplacian Eigenmaps* over the country-similarity graph G.

The feature that better describes the graph is the one that minimizes the difference between values along edges in the country-similarity graph. For each feature, we define a vector x, where x_i represents the corresponding feature of the country $c_i \in G$. As defined previously, w_{ij} represents the similarity between the countries c_i and c_j, so the idea is to find the feature x that minimizes: $LE(x) = \sum_{ij} w_{ij}(x_i - x_j)^2$

The latter can be calculated over the graph structure using matrix operations. We calculate the Laplacian of the graph, $L = D - A$, where A is the similarity matrix, and D is the diagonal matrix $diag(A \times 1_{n\times 1})$, i.e. $D_{ii} = \sum_j wij$. Then, $LE(x)$ can be determined by $LE(x) = x^T Lx$. In our analysis, each vector was standardized to a normal distribution with $\tilde{x} = 0$ and $\sigma = 1$.

When sorting the list of features by the value of their Laplacian, we have that the top features are (1) oil consumption, (2) latitude of the center of the country, (3) HDI, (4) number of mobile telephones, and (5) total area. This analysis shows that the features that better explains the graph are the ones that represent the current quality of life of the people (HDI), country's wealth (oil consumption and number of mobile phones), as well as some geographical features intrinsic to the country (latitude and area).

Continent similarity. Analogous to the similarity between countries, we calculated the similarity among continents, shown in the World map in Figure 9. We can observe a tightly related group of continents (Europe, North America, Africa, and Asia), meaning the users located in these continents visit similar domains, while Oceania and South America are farther apart from this cluster of continents.

8. Geographic Dynamics

To further understand how user behavior changes along time, we made a similar analysis using a sample query log of Yahoo! from the year 2005.

We observed that traffic to gTLDs was similar in both periods of time, as well as the location of the servers hosting these sites. When comparing the traffic among countries, we observed that in 2005 the countries with more diversified sources (largest source entropy) were located in Europe, while in 2008, Asian countries have a more diversified set of sources.

From the demographic analysis, we observed that in 2005, the features that better describe traffic similarity were latitude, HDI, unemployment rate, migration rate, and area of the country. This confirms our observations that countries in a similar latitude, or with similar human development have similar traffic.

9. Conclusions

In the present work, we analyzed the data from a large query log to describe the way the user behaves based on their location and the URL clicked. This study allowed us to identify a series of relevant findings.

By analyzing the characteristics of gTLDs, we observed that the `.com` domain has the largest share of traffic from several countries. Also, it is possible to observe that most of the hosts in gTLDs are located in the United States, with few exceptions, which are distributed among multiple countries.

There exists an inverse correlation between the incoming traffic entropy and internal referral ratio. This seems natural since when a ccTLD has a narrow set of incoming countries, the country itself is the most important source.

We observed that traffic among countries is concentrated in a few domains (e.g. `.uk` and `.ru`), which can be explained by linguistic and cultural factors. By aggregating the traffic to a continent level, we observed that `.uk` is the most common domain among continents. Users located in North America, Europe, Asia, and Africa visit a similar set of domains, while South America and Oceania have different behavior.

Our results show that language can be more important than geography when analyzing search similarity. In addition, what is considered as external to a country is not too dependent to the continent where the country is located, and depends more on factors such as language.

From the demographic analysis of the Web usage, it is possible to observe that countries in a similar geographic latitude, or with a similar human development index, tend to have similar traffic destinations.

On 2008, ICANN announced that they will liberalize the gTLDs over the next couple of years, so methods and studies about how gTLDs are used will become more important over the years. Also as future work we would like to repeat the experiments on samples in different years to look for trends over time.

References

[1] R. Baeza-Yates and A. Tiberi, “Extracting semantic relations from query logs,” in *KDD ’07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM Press, 2007, pp. 76–85.

[2] J. S. Brown and P. Duguid, *The Social Life of Information*. Harvard Business School Press, February 2002.

[3] R. Baeza-Yates and C. Castillo, “Relationship between web links and trade,” in *WWW ’06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2006, pp. 927–928. [Online]. Available: http://portal.acm.org/citation.cfm?id=1135777.1135948

[4] B. J. Jansen and A. Spink, “How are we searching the world wide web?: a comparison of nine search engine transaction logs,” *Inf. Process. Manage.*, vol. 42, no. 1, pp. 248–263, 2006.

[5] M. Sanderson and Y. Han, “Search words and geography,” in *GIR ’07: Proceedings of the 4th ACM workshop on Geographical information retrieval*. New York, NY, USA: ACM, 2007, pp. 13–14.

[6] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, “Analysis of geographic queries in a search engine log,” in *LOCWEB ’08: Proceedings of the first international workshop on Location and the web*. New York, NY, USA: ACM, 2008, pp. 49–56.

[7] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, “Spatial variation in search engine queries,” in *WWW ’08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 357–366.

[8] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li, “Detecting dominant locations from search queries,” in *SIGIR ’05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2005, pp. 424–431.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, “Graph structure in the web: Experiments and models,” in *Proceedings of the Ninth Conference on World Wide Web*. Amsterdam, Netherlands: ACM Press, May 2000, pp. 309–320. [Online]. Available: http://www9.org/w9cdrom/160/160.html

[10] R. Baeza-Yates, C. Castillo, and E. Efthimiadis, “Characterization of national web domains,” *ACM Transactions on Internet Technology*, vol. 7, no. 2, May 2007. [Online]. Available: http://dx.doi.org/10.1145/1239971.1239973

[11] K. Bharat, B. W. Chang, M. Henzinger, and M. Ruhl, “Who links to whom: Mining linkage between web sites,” in *International Conference on Data Mining (ICDM)*. San Jose, California, USA: IEEE CS, 2001, pp. 51–58. [Online]. Available: http://labs.google.com/papers/mininglinkage.html

[12] M. J. Silva, B. Martins, M. Chaves, A. P. Afonso, and N. Cardoso, “Adding geographic scopes to web resources,” *Computers, Environment and Urban Systems*, vol. 30, no. 4, pp. 378–399, July 2006. [Online]. Available: http://www.sciencedirect.com/science/article/B6V9K-4JW7WJG-3/2/a5a9dd75b3d1446beea70fef6b424681

[13] “IANA - Domain Name System Structure and Delegation (RFC 1591),” http://tools.ietf.org/html/rfc1591, 1994.

[14] “ISO - English country names and code elements (ISO 3166-1),” http://www.iso.org/iso/country_codes/iso_3166_code_lists/-english_country_names_and_code_elements.htm.

[15] “IPligence,” http://www.ipligence.com/, 2008.

Key references: [11], [3]

Country-level features. We used a set of 24 features extracted from World Factbook (`cia.gov/cia/publications/factbook/`) and Human Development Report 07/08 (`hdr.undp.org/en/`): Latitude (country’s center), Unemployment (%), Migration (%), Area (total land), Language (official or predominant), Median age, Literacy (age 15+ can read and write), Life expectancy (birth), Sex ratio (birth), Continent, Longitude (country’s center), GDP per capita (PPP), Exports, Imports, Telephones - mobile, Electricity consumption, Oil consumption, Internet hosts, Population, Internet users, Inflation, Telephone - main lines, GDP (PPP), and Human Development Index (HDI).