Toward Google Borders

Antoine Mazieres, Samuel Huron

Toward Google Borders

Antoine Mazières
INRA-SenS, LIAFA, Fabelier
Paris, France
antoine.mazieres@gmail.com

Samuel Huron
INRIA, IRI, Fabelier
Paris, France
samuel.huron@cybunk.com

Figure 1. The three most popular suggestions on Google Search for the query "Why society" represented on a world map (suggestions in red).

ABSTRACT

Query logs let by user on search-engines have helped create efficient tools for trend analysis, from commercial use to forecasting epidemics. In this paper, we propose a new method and system for cultural trends analysis based on Google autocomplete suggestions. We present *Zeitgeist Borders*, a toolkit enabling any user to collect and analyze associations between queries, suggestions and various regions of the world. We report unexpected observations about several behavioural and geographical trends along with promising uses.

ACM Classification Keywords

H.5 Group and Organization Interfaces: Web-based interaction

Author Keywords

Digital humanities. Digital Studies. Human Factors. Cultural Trends. Autocompletion. Suggestion. Web. Reflexivity.

Introduction

A great amount of information is produced by the use of web search-engines, allowing whoever possesses this data to know who searches for what, along with several characteristics about each user (location, time, device, etc). The analysis of these logs allows for improving the user experience by suggesting or automatically completing search query. For instance, a user of Google Deutschland's search engine who queries "how to cook" will learn that the most associated query with "how to" on `google.de` is "how to save a life".

WebSci'13, May 1 – May 5, 2013, Paris, France.
ACM 978-1-4503-1889-1.

Query suggestion has several purposes, among them less typing, catching mistakes or repeating frequent searches. Additionally, suggestion might induce serendipity through surprising or complementary propositions.

These suggestions reflect the most popular associations between words and letters. This happens within a specific group into which users are categorized, e.g. by country, interest, search history. It gives an immediate feedback to the user about the search of others users within this group, or subset. The user reflects his query to a greater number of similar queries that reveal a common behaviour among the specific set of users to which he is identified. This informations enables the user to reverse-engineer the behaviour of this subset, by revealing its most common query. For instance, the French media compared suggestions on Google France to other domains and noticed the specific suggestion of "juif" (the French word for *Jew*) for queries about public persons[1, 8]. Questionning suggestions reveals cultural facts about the users from which the suggestions are produced and the learning and sharing of this information is one of the possible uses of search engines.

In this paper we argue that multiplying and combining the possible sub-sets of users from which suggestions are made, enables unexpected observations that could lead to new insights in fields such as journalism, ethnography, sociology, political science, among others. After identifying the Top Level Domains (TLD, e.g. .fr, .com, .de, .in) as a sub-set used by Google for providing its suggestions, we created *Zeitgeist Borders* (ZB), an interactive visualization of all suggestions made to all sub-sets, allowing a greater capacity of inquiry for the user.

Previous work

Query completion and recommendation systems are a widely studied topic in several domains of computer science such as *database technology* [2, 7], *information retrieval* [17, 19, 21] and *Web* [22].

However, less research has been done about using this type of system to study cultural trends at a world scale. *Google Zeitgeist* [14] shows the most common requests over a one year period and provide an interface to explore this topic over a world map. *Google Trends* [13] is a webservice providing trends on a user defined query relative to the total search-volume distributed over the world. Baram-Tsabari et al. [4] studied the potentials and limitations of Google trends. Also, Ginsberg et al. [9] presented a method to analyse search query logs to track influenza-like illness arround the world, this technique being now provided as a service called Google Flue [11]. Specifically related to Google suggestions, Baker et al. [3] focus on racist, sexist or homophobic suggestions to argue that the Google Autocomplete algorithm can reinforce stereotypes.

A Reverse engineering approach to Google Autocomplete

Google Autocomplete (GA) [10] is the name of the suggestion system used by Google since 2008 and enhanced in 2010 by *Google Instant* [12] making suggestion available on-the-fly while the user is typing.

While several publications describe it [5, 15], the system internals are not available to the public. For this reason, we decided to use a reverse-engineering approach to understand if different suggestions are served to different users and to understand the criterias for those differences. A reverse engineering approach is defined, in this case, by the analysis of an existing obfuscated system to identify its behaviour and create representations of it in a simplified form for a purpose [6].

We listened to the network traffic of a computer while queries were being typed in the search box of a Chrome browser. This revealed an HTTP request[1] transmitting what is being typed. A response occurs, indicating the suggestions to be made (Fig. 2). The absolute URL of the HTTP request transmit the query along with 3 pieces of informations about the user: his browser, his query and which Google's domain, i.e. TLD, is being used. Top Level Domains (TLD) are assigned by an American NGO (ICANN) and most of them refer to a specific country, such as `.fr` refers to France, `.de` to Germany or `.in` to India.

We reproduced a same query over all supported TLDs supported by Google[2] referring to a country and observed differences among the responses. As a simple measure of these differences, we quantified the occurences of every first suggestion received by typing each letter of the Latin alphabet among all TLDs. On average, approximately 30% of TLDs (57 over 186), were served with the exact same first suggestion for each letter. This means that an average of 72 different first suggestions are distributed over the other 128 TLDs. For approximately 70% of Google's domains, a unique first suggestion was shared by only two of them[3]. While this statistical inquiry does not allow us to deduce that this ratio is true for all queries, it is sufficient to affirm that TLDs, and thereore countries, are a determinant aspect for distributing suggestions among users.

[1]For instance: GET `http://www.google.fr/complete/search?sugexp=chrome,mod=0&client=chrome&q=how+to` HTTP/1.1

[2]Listed in `http://www.google.com/supported_domains`

[3]Data and code available at `https://github.com/fabelier/Zeitgeist-Borders/tree/master/data`

```
["how to",
        ["http:\/\/www.jailbreakiphone4.fr\/",
        "how to make it in america",
        "how to become parisian in one hour",
        "how to save a life"],
        ["Jailbreak iPhone 4 | Jailbreak iPhone 4S,
        4.3.5, 4.3.4, 4.3.3, 4.3","","","",""],
        [],
        {"google:suggesttype":
                ["NAVIGATION","QUERY","QUERY","QUERY"],
                "google:suggestrelevance":
                                        [601,600,551,550],
                "google:verbatimrelevance":1300}]
```

Figure 2. Response to a HTTP request on GA

GA is a system thats return the top matching queries as suggestions to queries being typed by users. Through our analysis of this system, we have shed light on the fact that suggestions differ from one user to another, including on the base of the TLD requested. Consequently, TLDs are a significant sub-set for Google's suggestion system, and for every query made, users are shown the most common associated Google search queries from their country.

Tool design

We built a tool enabling users to map, worldwide the suggestions made by Google for a specific query. Through local suggestion, the user is given some knowledge about the other users with whom he is associated to. Given that GA associates users by the domain they use, those domains define the borders within which this knowledge is acquired about other users. Our tool allows one to abstract himself from these borders and acquire knowledge about any possible sub-set or combination of sub-sets of users.

Data is collected by the user defined query on every Google domains. We collect completions over 186 TLDs. Zero to four completions are received after each of our requests, as a list of string ordered by relevance. In order to scale the local ranking of suggestion to a world-wide view, we defined a score for each suggestion. A global ranking is built by making the sum of those scores for each similar suggestion over all TLDs. This score allow us to rank TLD suggestions at a world scale and let one knows wether suggestions are shared or not with other countries. Once collected, this data is stored and sent to the web interface.

By entering a query, the user get a list of all suggestions made by the search engine on all its domains, ranked by number of occurrences. As shown in Fig. 3, a mouse over a specific suggestion highlights the relevant countries on a world map. Also, a mouse over a specific country returns the suggestions made by the related domain.

Our software is called *Zeitgeist Borders* (ZB) and is released under the GPL [18].

Limitations Every query on ZB triggers as many queries on Google servers as there are TLDs. For example, six queries using our software causes over 1000 queries on GA's infrastructure. Tens of queries in a row cause the user's IP address

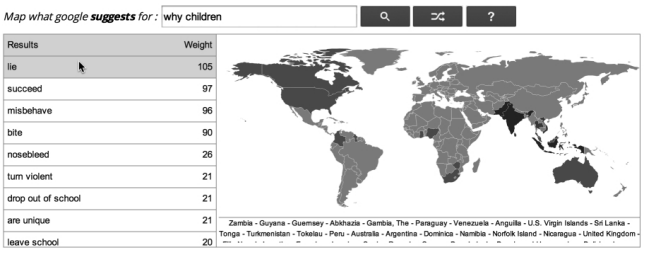

Figure 3. User interaction example on *Zeitgeist Borders*

to be banned from interacting with GA for a few hours. Such a limitation prevented us from collecting significant volumes of data for making statistical assumptions on the suggestion topology. However, with or without this limitation, important biases remain in the system:

- While Google is the most used search-engine on the web, its marketshare varies from one country to another. As an illustration, it has 97% of the market shares in Italy, Spain and Germany, while it is only below 40% of the market share in Japan, Russia and China [16]. This means that the use of Google search is in several countries limited to specific regions, users or topics.

- English appeared to be the language the most suitable for collecting suggestion from every domain. However, focusing on this language in a non-english speaking country constrains the returned suggestions to specific users (i.e. English-speakers) or matters (e.g. international news).

- GA's presence varies among domain and over time, sometimes along with Google search. This has been true during our research for Cuban and Chinese TLDs.

- According to GA policies [10], suggestions are turned off for certain sequences of characters associated with, for example, racism or pornographic topics.

Observations

As previously stated, our research does not pretend to shed light on the suggestion topology and distribution, since acquiring a volume of data that would be statistically relevant was not possible. However, we believe it is still worth reporting qualitative observations to build assumptions of the interests of the approach prototyped by our tool. Within these boundaries, we collected all suggestions of all TLDs for several list of queries:

- Letters of the Latin alphabet. This list was used in to assert the TLD-based system behind GA.

- Interrogative pronouns followed by most common verbs or pronouns: *what*, *where*, *when*, *how*, *who*, *why* followed by *is*, *are*, *was*, *I*, *he*, *she*. This list narrows suggestions to general-purpose interrogative queries.

- Intuitive queries according to our experience using ZB. These are mostly derived from the previous dataset with some added words making references to concept, name or thing. This non-structured list allows us to report unexpected observations without a specific methodology of inquiry.

With the Latin alphabet dataset, we noticed a strong presence of American or British companies. A measure revealed that 56% of the three most popular suggestions for each TLD among all letters are names of brands or products from USA or UK. For example, "amazon" is served as first suggestion among 40% of all TLDs over the letter "a". This means that the majority of suggestions over a single letter are linked to a few companies from only two countries in the world. Beyond highlighting an interesting economical and political fact, a generalized measure of this phenomenon would allow plotting of a semantic market share of those companies over their letter.

Interrogative pronouns are fundamental components of questions, and therefore a query with these is usually a formulated question. Adding pronouns or verb narrow the range to a category such as time (based on tense used) or gender ("he" or "she"). For instance, "what if" returns the hypothetical condition the most tested by Google's users. Whether from worry, fear or curiosity, users ask mostly about "if god was one of us", "if money didn't matter" and "if there was no google". The pronoun "why" asks for the meaning or cause of a phenomenon, the most common being "why is the sky blue". By adding a subject we can learn for what meaning is queried about a specific gender. For example, questions asked about the male gender include "why he disappeared" and "why he doesn't call", while those about the female gender asked "why she buys" and "why she slap". Also, different questions for different levels of languages can be identified by adding elements making a grammatically correct formulation. In that sense, being for she or he, "why does" returns "ignore me" and "love me".

Adding a concept to a pronoun-based query enable one to figure out what is mostly asked about this concept, or what other concept is linked to it. As shown in Figure 1, "why society" trigger a relation with complexity ("is a complex matter"), religion ("needs religion") and truth ("is wrong"). With names, ZB enables one to rank the popularity of a name and at the same time discover some elements of this popularity. In that sense, among all the Josephs, Joseph Gordon-Levitt is the more queried, this for his "girlfriend" and his role in "batman".

The visualization offered by our tool through a world map allow one to immediately test combinations or comparisons of group of users. For instance, the query "how to" clearly draw the north/south political separation over the world, with "how to kiss" for the south and "how to tie a tie" for the north. Also, different sub-groups can be tested for similarities. For instance, if communities of language seems also to provide similitude in queries, some sub-categories seems more tied together than others. For example, France and French-speaking African countries are frequently highlighted together while other French-speaking countries are not.

Last but not least, paying attention to uncommon suggestions allow identification of users with very specific preoccupations. For instance, while most of the countries ask if "why kids are the worst", or "lie", India is the only one asking why they "vomit" and "don't eat".

Discussion and future work

We have shown how monitoring results brought by search-engine suggestions can provide insights of cultural patterns across the world.

Our system is bound by the TLD-based Google suggestion system that was the only one available when conducting our research. TLDs are usually tied to political or administrative entities that do not necessarily match coherent cultural or even linguistic communities. Unfortunately, finer-grained information, such as individual search history or any data stored by Google other services, is not readily available. Obtaining richer information about users would allow aggregation of suggestions into a larger number and more diverse categories.

Moreover, a larger-scale analysis would allow systematic analysis of the cultural differences between countries. For example, two countries could be compared based on the Spearman coefficient [20] between exhaustive ordered sets of suggestions. We are currently working on a system enabling one to perform enough queries on GA's infrastructure to do so and to statistically ground any technical or social assertion.

Conclusion

This paper explored the possiblity of using Google Autocomplete (GA) suggestions for tracking cultural differences over the world. The tool, along with its method and system, enables users to collect, process, visualize and eventually analyze these suggestions.It also helps users to reveal how the group he belongs to biases the suggestions returned. We argue that autocompletion could provide a broad possibility for trend and social analysis. To that end, we reported several unexpected observations. By releasing our code as open source we hope to open doors and help future studies in this field.

Acknowledgments

We would like to thank very much Julien Palard for his great contributions to *Zeitgeist Borders* and to the vision it empowers. Also, we thank Jean-Philippe Cointet, Christophe Prieur, Joaquin Keller, Dusan Misevic, Camille Delebecque and Jagoda Walmy for their feedback and reviews on our research. Last but not least, we are very grateful to the communities of Fabelier[4] and the Center of Interdisciplinary Research[5] for the great environment they created.

REFERENCES

1. E. Anizon. Et françois hollande, il est juif ? *Telerama*, (3171), 06 2010.
2. R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology-EDBT 2004 Workshops*, pages 395–397. Springer, 2005.
3. P. Baker and A. Potts. Why do white people have thin lips? google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies*, 0(0):1–18, 0.
4. A. Baram-Tsabari and E. Segev. Exploring new web-based tools to identify public interest in science. *Public Understanding of Science*, 20(1):130–143, 2011.
5. P. Bialczak, W. Mazurczyk, and K. Szczypiorski. Sending hidden data via google suggest. *CoRR*, abs/1107.4062, 2011.
6. E. Chikofsky and I. Cross, J.H. Reverse engineering and design recovery: a taxonomy. *Software, IEEE*, 7(1):13 –17, jan. 1990.
7. A. Deshpande, Z. Ives, and V. Raman. Adaptive query processing. *Found. Trends databases*, 1(1):1–140, Jan. 2007.
8. S. Foucart. "juif", une requete très française. *Le Monde*, 02 2011.
9. J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
10. Google. Google autocomplete. `http://support.google.com/websearch/bin/answer.py?answer=106230`.
11. Google. Google flue. `http://www.google.org/flutrends`.
12. Google. Google instant. `http://www.google.fr/instant`.
13. Google. Google trends. `http://www.google.com/trends`.
14. Google. Google zeitgeist. `http://www.google.com/zeitgeist`.
15. G. R. Harik, S. Tong, and D. R. Cheng. Automatic completion of fragments of text. Patent 13/235,025, Google, 2012.
16. G. Light. Search engine marker share around the world - q4 2010. `http://www.greenlightdigital.com/assets/images/market-share-large.png`.
17. M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '97, pages 324–332, New York, NY, USA, 1997. ACM.
18. A. Mazieres, S. Huron, and J. Palard. Zeitgeist borders. `https://github.com/fabelier/Zeitgeist-Borders`.
19. Y. Nemeth, B. Shapira, and M. Taeib-Maimon. Evaluation of the real and perceived value of automatic and interactive query expansion. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 526–527, New York, NY, USA, 2004. ACM.
20. C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.
21. B. Tan, A. Velivelli, H. Fang, and C. Zhai. Term feedback for information retrieval with language models. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 263–270, New York, NY, USA, 2007. ACM.
22. Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 1039–1040, New York, NY, USA, 2006. ACM.

[4] `http://fabelier.org`

[5] `http://www.cri-paris.org/en`