The language specialisation of the Google search engine

Volker Schatz, September 2009

http://www.volkerschatz.com/science/nonpapers/

The most popular web search engine, Google, adapts its results and rankings according to the language setting of the client browser. This causes results to differ in ways intransparent to non-expert users, and leads to different views of the web being presented to different users. This study investigates this systematically. To that end, a novel method for comparing URL lists is developed, which separates differences in the URLs contained from differences in order and in rank. Google results were retrieved from different country-specific servers and by client computers located in different regions and countries, and were compared in view of language-related differences. The results of this study offer lessons about how to use Google in a research context, and how to conduct web searches with a reasonable standard of objectivity.

Keywords: **Google, web search engine, language**

I. INTRODUCTION

Google[BP98, PBMW99] has been the most popular web search engine for years.[com07, com09, Hit09] As such, it has a significant influence on the view of a large number of people experience of the web. It has also sometimes served as the basis of research not directly related to the World Wide Web.[CV07, EKH06] That prompts the question of how consistent and objective Google is, in other words whether it presents the same view of the world to all its users. Anecdotal evidence suggests that the number of hits reported by Google is inaccurate[Not03, EKH06] and can depend on the ISP used to connect to the internet.[EKH06]

But most systematically, Google takes the language setting of the browser sending a query into account. It returns a page in the appropriate language and redirects the request to a country-specific server google.*cc*, where *cc* is the country code top-level domain. This is documented in Google's FAQ.[Goo09] The work presented in this report investigates in how far Google's language specialisation affects its results.

II. METHOD

A. Overview of the investigation and data sets

As was mentioned in the Introduction, the topic under investigation is the dependency of Google web search results on the language preference of the user performing the search. In order to quantify that effect, using the search terms presented in the next section, three sets of result pages were retrieved to allow one-to-one comparisons.

The first two data sets were each retrieved within hours from a computer located in Germany. The first was intended for examining results from country-specific Google servers with different country-code top-level domains. Result pages of the same queries from different country-specific Google servers were retrieved. The second set of search results was obtained to verify that changes in URLs' Google ranking over time had not affected the validity of the analysis of the first. It was the result of an automated retrieval of results at the start and at the end of a time interval which retrieving the first data set took. This was repeated three times at intervals of one month.

The third data set, which was most elaborate to retrieve, was designed to investigate the effects of Google's language specialisation for users of the Windows operating system and Microsoft's browser, Internet Explorer, in different countries. The windows operating system is language-specific in that it does not allow the user to adjust its language after installation. Though the language of Internet Explorer can be changed, it is likely that most users are not aware of that fact, and still less aware that this setting may influence web search results. As a consequence, most end users are likely to use Internet Explorer "out of the box", with the default settings.

The third data set was obtained to put this hypothesis to the test. In a two-day tour, the same queries to google.com were performed from internet cafes in France, Switzerland and Austria, using the Internet Explorer browser provided without changing any of the settings.[1] On the third day, the queries were repeated from a private computer in Germany, under the same conditions.

For all data sets, only the first page of results was retrieved. This can be considered the most important portion of the results for search engine users, many of whom only consult the first page. However, results with both ten (the default) and 50 results per page were retrieved, and subsequently analysed. The inclusion of pages with 50 results was designed to show up differences in less highly ranked result URLs.

B. Choice of search terms

Performing comparisons of search engine results requires search terms. Here, terms were chosen which could plausibly appear as entries in an encyclopedia and about which one would reasonably expect to find relevant and objective information on the World Wide Web. It is for such terms that differences in results for different users can be considered most unwelcome. In addition, the search terms were selected to be largely language-neutral, that is, to be in use by speakers of

---

[1]Italy was also included in the tour, but due to a misconfiguration of the computers at the internet cafe in question, the results could only be saved in a proprietary printer format. This format could not be decoded, causing the data to be lost.

different languages. This was achieved by choosing technical terms and proper names.

Table 1 displays the six search terms. "ARGV" and "hard disk benchmark" are two computer-related terms. "ARGV" is the name of a program's argument array in the programming language C and others. "hard disk benchmark" is a technical term referring to a performance comparison between hard disks. Though an English expression, it is sufficiently well known among computer aficionados to be used also by non-native English speakers.

"Autocorrelation" and "laser" are scientific terms, referring to a signal processing method and a monochromatic coherent light source, respectively. Both exist in many European languages, with some differences in spelling for "autocorrelation". In the case of "laser", that is due to its being an acronym, in the case of "autocorrelation" due to its Latin roots.

The last two terms, "Shelley" and "Washington", both refer to persons. The former primarily denotes Percy Bysshe Shelley, a Romantic poet, or his wife Mary. The latter is the name of the first president of the United States and the American capital and an American state which were named after him.

C. Performing the queries

The different parts of this investigation treat the dependence of results on the country-specific Google server and on other factors separately. For that reason, it was necessary to send specific queries to specific servers of the Google search engine. This cannot be achieved using its web interface. In European countries, requests to `google.com` are redirected to the appropriate country-specific Google web page. In addition, queries performed using the web interface often have additional query parameters appended, which prevents one-to-one comparisons.

For these reasons, all requests to the Google search engine were performed by typing the query URL directly into the location bar of the browser. Query URLs have the following form:

`http://www.google.cc/search?q=term`

Here *cc* was replaced by `com` for the American Google server, `co.uk` for its British server and the respective two-letter

Term	Explanation
ARGV	Program argument list
Autocorrelation	Signal processing method
Hard disk benchmark	Test method or result for hard disks
Laser	Coherent light source
Shelley	Romantic poet
Washington	Capital, state and first president of the US

Table 1. Search terms used for the comparisons, and brief explanations. All search terms were used in lower case; they are capitalised here only for reasons of presentation.

country code for the others. *term* stands for the search term. All search terms were typed in lower case. In multiple-word queries, the words are separated by plus signs. The only term used here for which this is applicable is "hard disk benchmark". For those requests in which 50 rather than 10 search results were obtained, the string "`&num=50`" was appended to the URL given above. Giving this number explicitly was not necessary (and was not done) for requesting ten results, as this is the default.

D. Extraction of result URLs

The query results were extracted from the retrieved HTML documents by automated "screen scraping". A Perl script was written to parse the documents and print out the list of result URLs. This was a non-trivial task: These documents contain hyperlinks other than the search results, and some of them are hyperlinked to offers of translations and maps from Google. The result URLs could be distinguished by the fact that they were both hyperlinked and printed as text, in whole or in part, in the HTML document below the hyperlink. This fact, as well as discarding hyperlinks to servers such as `translate.google.com` and `maps.google.com`, allowed to extract the query results exclusively in most cases.

The correctness of the resulting URL lists was verified in two ways. That the number of extracted results was correct was checked in an automated way, using a complex UNIX shell command to print all result files containing more or fewer lines than they should have. Differences from the expected number were rectified by tweaking the scraping script or by manual correction. Secondly, the correctness of the result URLs themselves was verified visually for random samples.

E. Comparison of URL lists

There is a sizeable body of work about the evaluation and comparison of search engine results, see e.g. [HCBG01, BI02, SL02] and references therein. However, due to its origins in the evaluation of information retrieval systems, most of that work aims at measuring performance in absolute terms, with the absolute results then being compared. This traditionally has often involved using human judges. The effort and difficulties involved in this approach are unnecessary in this investigation, the purpose of which is to detect differences in the list of result URLs rather than assess absolute quality. Therefore, in our case it is quite sufficient to compare corresponding lists of result URLs.

Bar-Ilan[BI02] has also compared the sets of URLs returned by search engines, though with a different intent, namely to detect changes in results over time. To that end, she records all hits returned by a search engine and tracks changes of that set. In more recent work,[FKS03, BIMHL06] comparison methods which treat web search results as ordered lists have been investigated and used. The approach used here follows them in spirit if not in detail.

Two lists of URLs can differ in two respects: the URLs they contain and the order in which they occur (their ranking). To catch all aspects of differences between URL lists, three different measures will be used. They will be briefly and

largely verbally described in the following; a more precise and mathematical definition can be found in the Appendix.

The first measure of difference is the proportion of URLs present in one list but not in the other. It lies between 0 and 1 and is 0 for identical lists. This is both the most obvious and most important measure of the differences between search result lists.

The other two measures reflect the ranking of the search results. They are computed from the ranks of those URLs which are common to both lists. The first of these two measures indicates how many URLs have their relative rank reversed in one list relative to the other. This expresses how different the importance of the common URLs is considered relative to each other in both lists. It is computed by counting all pairs of URLs which occur in opposite order in the two lists. Mathematically, this is the number of inversions of the permutation mapping the common URLs' order in the first list to that in the second. This number is normalised by dividing it by the maximum possible number of such inversions, $n(n-1)/2$, where n is the number of common URLs. This again yields a number between 0 and 1, which is 0 for identical lists or lists in which the common URLs are in the same order.

The third measure quantifies how different the rank of the same URL is in the two lists even when common URLs are in the same order. It is designed to indicate situations such as when the same set of URLs appears near the top in one list, and in the same order but near the bottom in the other list. To compute it, the differences in rank are averaged for all common URLs and then normalised by dividing the average by the maximum difference, the total number of URLs in the lists minus the number of common URLs. The result is also between 0 and 1, and 0 for equal lists.

To sum up, three measures of the difference between lists of URLs have been defined. All are normalised to lie in the range between 0 and 1, and are 0 for identical lists. In the following text, the differences quantified by them will be called differences in result URLs, order and rank, respectively.

After comparing individual URL lists as just described, the three difference measures were averaged over all variables except the one with respect to which the comparison was performed. For instance, the data analysed in Section III-A were compared with respect to the country-specific Google server from which they were obtained. The results displayed in Table 2 were obtained by computing the difference measures between URL result lists from different servers but the same search term and result count, and then averaging each difference measure over the search term and result count.

F. Classification of results according to language and TLD

Besides comparing lists of search results, the URLs were also classified according to their language and the top-level domain (TLD) of their servers. The languages of the result pages were determined manually. Mixed-language pages were counted towards the non-English results. Pages with navigation items in a different language from the main content were not classified as mixed-language, but those with forum posts in different languages were. The breakdown into minority languages in the results sections always covers the complete set of result URLs, i.e. all other results were in the English language.

In order to tally the URL results for each TLD, a small Perl script was created which splits each URL up, extracts the TLD from the server name, and counts the URLs with a given TLD.

III. Results

A. Comparison by country-specific server

Queries to Google servers with different top-level domains are where differences in results are most obviously to be expected and most transparent to the user. This effect was quantified in the first part of this investigation, the results of which are presented in Table 2. The share of result URLs which differ is between 9 % and 24 %. The common URLs are shuffled to a modest extent. However, the similarity of results between countries which share the same language, such as .com (America) and .uk, and .at, .ch[2] and .de, is not systematically larger than between countries with different languages. The largest difference in results is between google.ch and google.de, the smallest between google.com, google.fr and google.it (equal for all three pairwise comparisons).

The vast majority of results from all servers are English-language web pages. The frequency of occurrence of web pages in other languages is shown in Table 3. German is the most frequent minority language in this set of results, accounting for more than a third of results, followed by Italian. The variation in the share of each language depending on which server was queried is small. google.de, google.fr, google.it and google.co.uk returned slightly more results in their native language. (In the case of google.co.uk, this is visible in the table by the reduced share of German.) But the distribution of languages of results returned by the other three servers is almost identical. A detailed inspection showed that this is also true for each search term separately.

Only a minority of the result URLs referred to websites with the same top-level domain as the Google server which returned them. A larger proportion of results referred to sites at the .de top-level domain, as shown in Table 3. Together with the large share of German-language results, this suggests that these results were tailored towards consumption in Germany, where the client computer performing the requests was located. This observation gave rise to the investigation presented in Section III-C.

B. Changes of results over time

The results presented above are based on retrieving sets of considerable numbers of Google result pages. As described in Section II-C, these pages were retrieved manually, using a web browser. As a consequence, it took several hours to retrieve the

---

[2]Though German is only one of four languages spoken in Switzerland, it is the most prevalent.

Server TLD	.at	.ch	.com	.de	.fr	.it
.uk	0.16 0.03 0.14	0.18 0.11 0.26	0.11 0.03 0.17	0.20 0.08 0.18	0.12 0.02 0.16	0.12 0.03 0.19
.it	0.15 0.04 0.19	0.19 0.10 0.25	0.09 0.04 0.16	0.20 0.10 0.10	0.09 0.02 0.13	
.fr	0.12 0.02 0.08	0.16 0.08 0.24	0.09 0.03 0.07	0.19 0.07 0.11		
.de	0.21 0.08 0.14	0.24 0.14 0.21	0.13 0.07 0.09			
.com	0.13 0.05 0.12	0.17 0.11 0.26				
.ch	0.19 0.07 0.25					

Table 2. Differences in results returned by the Google server with the given top-level domain (TLD). The three numbers for each comparison represent differences in results, order and rank as described in Section II-E.

Server	German	Italian	Spanish	French	Polish	server TLD	TLD .de
google.at	141	10	4	2	0	37	75
google.ch	145	8	4	2	0	53	65
google.com	141	10	4	3	0	96	100
google.de	170	8	3	1	0	135	135
google.fr	128	10	4	19	1	8	79
google.it	125	31	4	3	1	18	81
google.co.uk	129	9	4	3	1	24	88

Table 3. Minority languages and top-level domains (TLDs) of web pages by Google server returning the result. The first numerical columns present the number of web pages in languages other than English returned by the given Google server. The last two columns give the number of results with the same TLD as the Google server and with the TLD .de, respectively. The total number of results from each server, which these numbers relate to, was 360, one page with ten results and one with 50 results for each search term.

Client location	A	CH (de)	CH (it)	D
F	0.54 0.07 0.15	0.53 0.07 0.15	0.53 0.24 0.23	0.54 0.07 0.15
D	0.04 0.01 0.14	0.05 0.03 0.21	0.50 0.10 0.23	
CH (it)	0.50 0.14 0.24	0.51 0.14 0.22		
CH (de)	0.04 0.03 0.09			

Table 4. Differences in results returned by google.com when queried from different countries and regions. The three numbers for each comparison represent differences in results, order and rank as described in Section II-E. For Switzerland, the country code corresponding to the local language is given in parentheses. The differences between locations which share the same language are strikingly smaller than between places with a different local language.

Client location	German	French	Italian	Spanish	local TLD	local language TLDs
A	111	1	0	0	1	85
CH (de)	111	1	0	0	8	87
CH (it)	0	1	89	1	0	52
D	115	1	0	0	79	88
F	0	124	0	0	41	41

Table 5. Minority languages and top-level domains (TLDs) of web pages by client location. The first numerical columns present the number of web pages in languages other than English returned by google.com when queried from client computers at different locations. The last but one column gives the number of results with the TLD corresponding to the client location. The last column contains the total of results with TLDs corresponding to countries in which the same language is spoken. For the purpose of the last column, the .ch TLD was counted as German-speaking, since most Swiss web pages are in German. The total number of results from each server, which these numbers relate to, was 360, one page with ten results and one with 50 results for each search term.

complete set of result pages. The differences found may thus have been influenced by changes in the result URLs' Google ranking over time.

To guard against such a confounding influence, a further investigation was performed. Three times over the duration of two months, sets of ten and 50 results were retrieved from `google.com` both before and after a $3\frac{1}{2}$-hour interval. This interval approximately equals the time it took to retrieve the complete first data set manually. So as to retrieve the result pages for this investigation within as short a time as possible, they were retrieved in an automated fashion using the browser Links. This browser can be used as a text-only or graphical browser as well as an automated downloader, and is not blocked as a robot by the Google servers.

Corresponding search results from before and after each $3\frac{1}{2}$-hour interval were compared, and the differences averaged over the search terms, number of requested results and the three occasions on which these result pages were obtained. The result is a 7% change in result URLs, a 2% difference in their order and a 7% difference in rank. These values can serve as a baseline for the results of the previous section. The differences found in Section III-A indeed exceed that baseline, except for differences in order and rank in a few cases. They are therefore not just a statistical fluke due to changes in the results' Google rankings.

C. Comparison by client location

This section presents comparisons of Google results between client computers located in different countries and regions with different languages. The results are shown in Table 4 and 5. The differences are striking: The three German-speaking locations — Austria, German-speaking Switzerland and Germany — show differences in results of 5% or less, with some modest differences in order and rank. By contrast, differences between locations where different languages are spoken are at least 50%.

A breakdown of results by top-level domain and language confirms this picture. The classification by language was performed manually, as described in Section II-F. As can be seen in Table 5, the local language was second only to English regarding the number of result pages. The same is true for the number of results with top-level domains which correspond to the local language. These results are consistent with Table 3, where results in the German language and with the top-level domain `.de` were found to be prevalent among results retrieved from Germany.

The reason for these differences is revealed by the HTTP headers of the requests, which were obtained from a CGI script installed on the author's web server[3]. The value of the HTTP_ACCEPT_LANGUAGE header was correctly set to the local language and is the probable reason for the differences in results. It should be noted that the HTTP_ACCEPT_LANGUAGE header is apparently much more influential than the TLD of the server queried, as the

much larger differences in Table 4 compared to Table 2 show. This is important, as the HTTP headers are far less transparent to the user than the TLD of the server, which is displayed in the browser location bar.

As was mentioned in Section II-A, further results were retrieved from Italy, but could not be saved in HTML format due to a misconfiguration in the internet cafe in question. The impression gained while retrieving these search results was that they contained many Italian-language pages, thus anecdotally confirming the findings above.

IV. DISCUSSION

The results presented in the previous sections show that the results obtained with the web search engine Google depend significantly on which country-specific Google server is queried and on the language settings of the browser. In particular, a user who queries Google with a stock Windows system in its default configuration will be presented search results heavily tailored towards the locally prevalent language, even when the search terms are language-neutral.

This language specialisation is apparently part of the service Google provides for its customers, and meets the needs of most. However, it is of some concern when Google is used for research purposes, whether related to information retrieval or not. Scientific work should be reproducible, but unless care is taken, data obtained via Google may be skewed towards the researcher's country and language. At a minimum, the HTTP_ACCEPT_LANGUAGE header sent by the browser used should be included in publications. Preferably, queries should be repeated with different browser language settings, and different country domain servers should be queried by explicit URL. The results should then be compared for systematic differences relevant to the research topic under investigation. The same precautions should be taken when using other search engines, which can be expected to include similar features.

Considering the effort involved in that strategy, it would be preferable to avoid the language bias altogether. Investigating whether search engine APIs also exhibit language specialisation is beyond the scope of this work, but there are grounds for hope that they do not. The APIs are aimed at researchers rather than consumers. And while the protocols used are mostly HTTP-based, the request is embedded in the payload rather than the URL and is therefore likely to be processed separately from the HTTP headers. On the downside, the APIs suffer from reliability problems and volume restrictions[MT05a, MT05b] and seem to be based on smaller indices.[MN07, TM09] Google's new AJAX API also places severe restrictions on its users.[TM09]

For those reasons, scraping results from the results of web search interfaces is likely to retain its place in the repertoire of research methods. Researchers should be aware of the language specialisation feature and observe the rule of *caveat user*.

APPENDIX

This Appendix will give a more formal and precise definition of our tripartite measure of similarity for URL lists

---

[3]The script in question is a Perl script which displays all available environment variables, including the HTTP headers. It can be obtained from http://de.selfhtml.org/servercgi/cgi/umgebungsvariablen.htm.

than the verbal explanation in Section II-E. For generality, it will not be assumed that the lists to be compared have the same size. Lists of different lengths will be regarded as equal if the longer list has the shorter list as its prefix, in other words, if it extends it by appending further URLs but does not insert others in between the leading entries.

The set of all URLs shall be denoted by U, elements of which shall be represented by small letters, such as for example u and v. Obviously no computations can be performed on the set of URLs, but they can be compared for equality and inequality.

Lists of URLs will be denoted similarly to vectors in mathematics to express the fact that the order of the URLs contained matters:

$$L = (u_1, u_2, \ldots, u_N), \qquad u_i \in U\,.$$

We denote the element of a list at a given position with square brackets, for example for the list defined above:

$$L[i] = u_i \qquad \text{for} \;\; i = 1 \ldots N\,.$$

As the URL lists we are dealing with are ranked search results, all the entries of lists used below will be different:

$$L[i] \neq L[j] \quad \forall\, i \neq j$$

Under this condition, the rank of a URL that is part of a list can be defined as

$$R(L[i], L) := i\,.$$

When the order of the URLs is of no importance, the set of the URLs in a list shall be written briefly as

$$\{L\} := \{u_1, u_2, \ldots, u_N\}\,.$$

The size of a list, that is, the number of URLs it contains, will be denoted by

$$\#L := \#\{L\}\,,$$

which is equal to N for the example list above.

The most obviously important criterion for comparing two URL lists is the share of URLs they have in common. In order to write this down formally, it is expedient to define a Kronecker delta for URLs.

$$\delta_{uv} = \begin{cases} 1 & \text{for } u = v \\ 0 & \text{otherwise} \end{cases}$$

Then the number of URLs that two lists L_1 and L_2 have in common can be computed as follows.

$$C(L_1, L_2) := \sum_{u \in \{L_1\},\; v \in \{L_2\}} \delta_{uv} \qquad (1)$$

This makes use of the fact that the lists have no duplicate entries and that therefore a given entry of one list can equal at most one in the other list. As a consequence, $C(L_1, L_2)$ can be at least zero and at most the number of URLs of the shorter list. By normalising this quantity, one can define a number $d(L_1, L_2) \in [0, 1]$ representing how different the contents of two lists are.

$$d(L_1, L_2) := 1 - \frac{1}{\min(\#L_1, \#L_2)} C(L_1, L_2) \qquad (2)$$

$d(L_1, L_2)$ takes the value 0 for identical lists, and 1 for disjoint lists. If L_1 and L_2 have different sizes, and one is contained in the other, it also takes the value 0. This reflects the above prescription that a longer list should be viewed as consistent with a shorter list if it contains it as a prefix.

The measure (2) is sufficient for comparing the sets of result URLs returned by two search engine queries, but does not account for differences in the ranking of the results they have in common. Differences in ranks of the same URLs shall be analysed separately in two respects: the order of the common entries among themselves, and their ranks in the complete list.

The first is well described by a permutation which maps the common entries in the order in which they appear in the first list to the order in which they appear in the second. This comparison will ignore the URLs which do not occur in both lists. Therefore it can be viewed as a comparison of only the partial lists which result from our original lists by removal of the URLs which are not contained in both. These partial lists have the size $C(L_1, L_2)$ and will be denoted by L_1' and L_2'. After numbering the URLs in the order in which they appear in L_1', the permutation π is defined as follows:

$$L_1' = (u_1, u_2, \ldots, u_{C(L_1, L_2)})$$
$$\pi \in \Pi_{C(L_1, L_2)} \qquad (3)$$
$$\text{such that } L_2' = (u_{\pi_1}, u_{\pi_2}, \ldots, u_{\pi_{C(L_1, L_2)}})$$

Here Π_n denotes the set of permutations of n elements.

In order to describe the different order of URLs in L_1' and L_2', the notion of an inversion of a permutation is needed. An inversion is a pair of elements whose order is exchanged by the permutation:

$$i, j \in \{1, \ldots, C(L_1, L_2)\} \text{ such that } i < j \;\wedge\; \pi_i > \pi_j$$

The number of inversions of the permutation π is a measure of how badly the order of the common URLs has been shaken up. It can be written as follows.

$$I(L_1', L_2') := \sum_{i, j \in \{1,\,\ldots,\,C(L_1, L_2)\},\; i < j} \begin{cases} 1 & \text{for } \pi_i > \pi_j \\ 0 & \text{otherwise} \end{cases}$$

In order to obtain a quantity whose range of values does not depend on the size of the URL lists, this expression is normalised:

$$\iota(L_1, L_2) := \frac{I(L_1', L_2')}{C(L_1, L_2)\big(C(L_1, L_2) - 1\big)/2} \qquad (4)$$

$\iota(L_1, L_2)$ takes the value 0 for identical lists or lists which contain their common URLs in the same order. It takes its maximum value 1 for lists in which the order of the common URLs is reversed. In the special case where L_1 and L_2 have no or one URL in common, ι is also defined as zero.

$$\iota(L_1, L_2) := 0 \quad \text{for} \;\; C(L_1, L_2) \in \{0, 1\} \qquad (5)$$

In d and ι we have two quantities reflecting how many URLs are common to two lists and how the order of the common URLs differs. What neither of them catches is differences in the absolute ranks of the common URLs within the complete lists L_1 and L_2. This especially concerns the comparison of lists of different lengths, which we have decided should be

equal if the longer starts with the entries of the shorter, but not if it contains them further down, even in the same order.

To account for such differences, a further quantity will be defined. Since ι already describes rearrangements of the common URLs among themselves, it makes sense to make that third quantity independent of the order of the common URLs. We define it as a suitably normalised average of the differences in rank of the common URLs, without regard to their order:

$$\varrho(L_1, L_2) := \frac{1}{C(L_1, L_2)\,\big(\max\{\#L_1, \#L_2\} - C(L_1, L_2)\big)} \cdot$$
$$\cdot \sum_{i\in\{1\ldots C(L_1,L_2)\}} \big| R(L_1'[i], L_1) - R(L_2'[i], L_2) \big|$$
$$\text{for } 0 < C(L_1, L_2) < \max\{\#L_1, \#L_2\}. \quad (6)$$

Note that the difference of ranks is formed not between the same URLs, but between URLs with the same rank in L_1' and L_2', so as to achieve independence of the order of the common URLs. The division by $C(L_1, L_2)$ arises from the average over the common URLs, and $(\max\{\#L_1, \#L_2\} - C(L_1, L_2))$ is the maximum possible difference in the average rank, division by which normalises the result. For completeness, ϱ is defined as zero for both disjoint and identical lists:

$$\begin{aligned} \varrho(L_1, L_2) := 0 \quad &\text{for } C(L_1, L_2) = 0 \\ &\text{or } C(L_1, L_2) = \#L_1 = \#L_2 \end{aligned} \quad (7)$$

As for d and ι, the values of ϱ lie in the interval $[0, 1]$. ϱ is 0 for identical lists and 1 when the common URLs are at the top of one list and at the bottom of the other.

The three measures d, ι and ϱ are independent and complete in the following sense: Regardless of the value of one of them, the others can still take an arbitrary value, except in the slightly pathological cases of the two lists having no or just one URL in common. And there are no two non-identical lists which do not cause at least one of the measures to be non-zero.

The first statement, of independence, has to be qualified with the fact that the set of values ι and ϱ can take depends on the number of common URLs. This also gives rise to the two pathologies for only one common URL (where ι is always 0) or disjoint lists (where both ι and ϱ are 0). Under normal circumstances, however, ι and ϱ can take the values 0, 1 and any discrete intermediate value arising from a possible permutation of the common URLs. ι and ϱ are also independent of each other, as ι depends only on the order of the common URLs among themselves and ϱ is by construction independent of that order.

The property of completeness is equally easy to see. If the two lists of the same length do not have all elements in common, $d > 0$. Otherwise there is a permutation mapping the complete list L_1 to L_2. The only permutation which has no inversions and for which ι would consequently be 0 is the identity, which contradicts our assumption of having non-identical lists. So at least one of d or ι is non-zero for non-identical lists of equal length. For lists of different lengths, $d > 0$ if the longer list does not contain all elements of the shorter. Otherwise, $\iota > 0$ unless the longer list contains the

elements of the shorter in the same order. If both d and ι are zero, ϱ is still non-zero unless the shorter list is a prefix of the longer. But this is the situation which was defined as equality of differently-sized lists at the start of this Appendix. So at least one of d, ι and ϱ is non-zero unless the lists compared are equal.

REFERENCES

∴ prior work   ⇔ related   ◁ background

[BI02]   ∴ J. Bar-Ilan: *Methods for Measuring Search Engine Performance over Time*. Journal of the American Society for Information Science and Technology **53** (4) (2002) 308–319. doi:10.1002/asi.10047.

[BIMHL06] ∴ J. Bar-Ilan, M. Mat-Hassan & M. Levene: *Methods for comparing rankings of search engine results*. Computer Networks **50** (10) (2006) 1448–1463. Online.

[BP98]   ◁ S. Brin & L. Page: *The anatomy of a large-scale hypertextual Web search engine*. In Proceedings of the seventh international conference on World Wide Web, vol. 30. April 1998 pp. 107–117. doi:10.1016/S0169-7552(98)00110-X. Online.

[com07] ◁ comScore Inc.: *Worldwide search market report*. Press release. October 2007. Online. Archived.

[com09] ◁ comScore Inc.: *March 2009 U.S. Search Engine Rankings*. Press release. April 2009. Online. Archived.

[CV07]   ⇔ R. L. Cilibrasi & P. M. Vitányi: *The Google Similarity Distance*. IEEE Trans. Knowledge and Data Engineering **19** (3) (2007) 370–383. Online. ArXiv:cs/0412098v3 [cs.CL].

[EKH06] ⇔ A. J. Evangelista & B. rn Kjos-Hanssen: *Google distance between words*. Presented at Frontiers in Undergraduate Research, 2006. 2006. Online. ArXiv:0901.4180v1 [cs.CL].

[FKS03] ∴ R. Fagin, R. Kumar & D. Sivakumar: *Comparing top k lists*. SIAM Journal on Discrete Mathematics **17** (1) (2003) 134–160.

[Goo09] ◁ Google Inc.: *Google General FAQ*. Web page. January 2009. Online. Archived.

[HCBG01] ⇔ D. Hawking, N. Craswell, P. Bailey & K. Griffiths: *Measuring Search Engine Quality*. Information Retrieval **4** (1) (2001) 33–59.

[Hit09] ◁ Hitwise corporation: *Google Hovering at 72 Percent of U.S. Searches in March 2009*. Press release. April 2009. Online. Archived.

[MN07]   ⇔ F. McCown & M. L. Nelson: *Agreeing to Disagree: Search Engines and Their Public Interfaces*. In JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries. ISBN 978-1-59593-644-8. 2007 pp. 309–318. doi:10.1145/1255175.1255237. Online.

[MT05a] ⇔ P. Mayr & F. Tosques: *Google Web APIs - An instrument for webometric analyses?* In P. Ingwersen & B. Larsen (eds.), Proceedings of the ISSI 2005 conference. 2005 pp. 677–678. Online. ArXiv:cs/0601103v1 [cs.IR].

[MT05b] ⇔ P. Mayr & F. Tosques: *Webometrische Analysen mit Hilfe der Google Web APIs*. Information: Wissenschaft und Praxis **56** (1) (2005) 41–48. Online. Archived.

[Not03] ∴ G. R. Notess: *Google Inconsistencies*. Web page. 2003. Online. Archived. Retrieved June 2009.

[PBMW99] ◁ L. Page, S. Brin, R. Motwani & T. Winograd: *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66, Stanford InfoLab. November 1999. Online.

[SL02]   ⇔ Y. Shang & L. Li: *Precision Evaluation of Search Engines*. World Wide Web **5** (2) (2002) 159–173.

[TM09]   ⇔ F. Tosques & P. Mayr: *Programmierschnittstellen der kommerziellen Suchmaschinen*, pp. 116–147. Akademische Verlagsgesellschaft AKA. 2009. Online.