The Effects of Personalization on Search Results

David Emmel- demmel3@gatech.edu
Daniel Kester- dkester3@gatech.edu
Yash Thadhani- ythadhani3@gatech.edu

Abstract

Our project was to study personalized web search results and how they change over time. To do this, we created a bot using Selenium that could record the results of a Google Search over time. We then had multiple users run the bot to record a number of different searches both when they are logged in and when they are not and evaluate how relevant the search results are over a period of time. By comparing the results using our total relevance metric, we found a slight difference between the personalized and standard results but no difference between the rates at which they change.

Motivation

Search engines like Google are some of the most important applications on the Internet. People rely on search engines to direct them to the best pages for the topics they search and websites rely on them to direct traffic to the website. One method search engines have used to try to improve the relevance of results is by personalizing them to specific users based on data like their past search history. Since search engines play such an important part of a person's Internet experience, it is worth examining just how effective these personalized searches are.

Objectives

- Are personalized search results more relevant than generic search results?
- Are personalized search results more relevant for topics that are not commonly searched by users?
- Do personalized search results become more relevant over time?
- Do personalized search results become more relevant faster than the standard search results?

Methodology

Our project was a three part process. First, we had to create a bot that could record search results. Next we had to have users evaluate the relevance of the result. Finally, we had to analyze these results over time. The details of each of these steps are described below.

In order to observe the change of searches over time, we created a bot using Selenium Web Driver. This bot would ask for a search query and if the user wanted a personalized result or a standard result. After that information was added, it would open Firefox and go to Google. If the user wanted personalized results, it would then go to the login page and allow the user to login. The bot itself would wait until Firefox was back on the Google home page since the user may have two factor authentication activated (so it couldn't submit immediately after the first login page). Once it was on the homepage (either after logging in for personalized or opening the browser for nonpersonalized), it would enter the search query into Google and begin the search. It would then go through the first 5 pages of results and record all the standard results (ignoring results like advertisements or "In the news" sections) into a CSV file with the URL's rank. The rank was how high the URL was in the results, calculated as (Total Number of URLs - (index -1)). Thus if there were 46 results, the first result was rank 46, the second was 45 and so on until the last result on page 5 which was rank 1.

Recording the results into the CSV was a multi step process. First, the bot had to see if a CSV for that search existed already or not. If so, it would read in the existing results and if not, it would create a new file. Next, it would look through any existing results and see if the URL matched that of any of our new results. If so, the rank of that URL for the new search was tacked on to its next column. If not, a new row was created, recording the url (to use as an ID in the future), the title (to help users determine the page's relevance), a rank of 0 for all previous searches that this url wasn't a part of and the rank of its current search. Finally, it stored a rank of 0 for all existing URLs that weren't part of the current search result.

At this point, we told the users who ran the searches to add the URL relevance of each result to the CSV file. The only information we gave about the sites was the URL and title recorded in the file (as this is what the user would see if they performed the search normally). The users were told to give each result a score from 1 to 5, with 5 being the most relevant and 1 being the least. The users were given no guidance as to what was considered "relevant" as if they had performed the searches themselves, they would just choose which result is best on their own.

Once all this information was recorded in the CSV, we could begin to compare the results across different CSVs. To do this, we created a metric that we called total relevance.

$$TotalRelevance = \frac{\sum_i (UrlRelevance(i) * Rank(i))}{5 * \sum_i (Rank(i))}$$

In this metric, UrlRelevance(i) is the score the relevance score the user gave a web page between 1 and 5 and Rank(i) is the rank the bot recorded for the page. We chose this metric over existing search metrics for a few reasons. First, many search metrics such as recall and precision require knowledge of all possible results, something we simply do not have. Second, the other metrics we found online did not allow us to both order our results and weight them (which we needed to consider both URL relevance and result rank). Finally, this metric allows us to compare against searches that have a different number of results (since it normalizes by dividing by the max possible score for a result). While our metric is likely not as sound as some of the others, it does accurately measure what we are looking for by giving more weight to results that are both encountered early on and are more relevant.

Challenges

There are many different challenges we ran into while working on this project, both technical and psychological. These challenges are described below.

The first major challenge was dealing with the asynchronous loading of search results in Google. Selenium normally will try to perform an action by using class and ids within the HTML on the page. If the element it is looking for is not there, it will simply crash. Because Google uses a single page for all its results and just loads them in asynchronously, by default we would have no way to see the results as it would crash while they were loading. To get around this problem, we simply have a loop that waits for the "Page X" or "About" at the top of the results to load before trying to read the results.

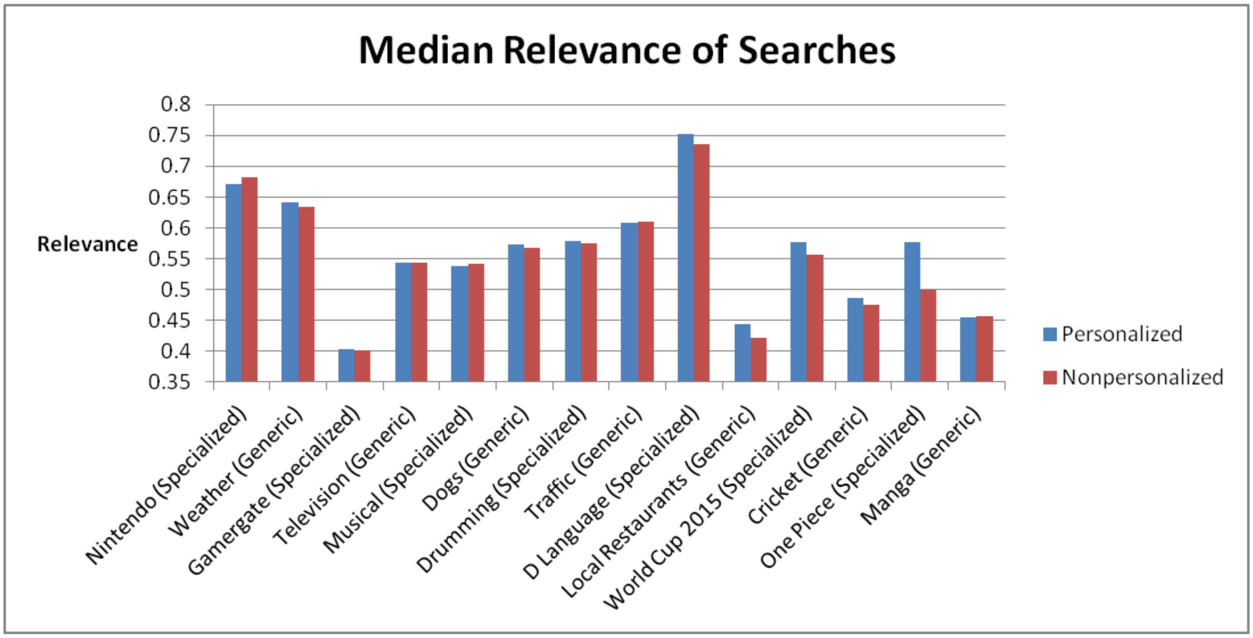
The next challenge was to deal with duplicate URLs. Occasionally, we would encounter URLs that show up on multiple pages in the Google search results. The way our bot works by default, it records both the results in separate columns in the CSV, which would mess up future searches. While we could have gotten around this by adding a check into the CSV code, we instead decided to manually remove the duplicate from the CSV. We felt that all that really mattered was the higher ranked result (as people would have clicked it on the first encounter if they were interested). Since our total relevance metric is normalized against the highest possible score with the given ranks, this wouldn't affect the results and would give us a more accurate representation of search relevance.

The third challenge was accounting for the location of the searches. As mentioned in class during our presentations, Google does some personalization based on the location of the IP address that performed the search and could potentially affect our results. While this did not matter for most of the searches (since the two that were being compared were run in the same location), it could affect the ones that were run in both Atlanta and Alpharetta. We decided instead of fixing this problem, we could measure it and compare the results. To do this, we manually marked which of the searches were performed in Alpharetta by adding an A to their date in the CSV files, allowing us to compare all the different results.

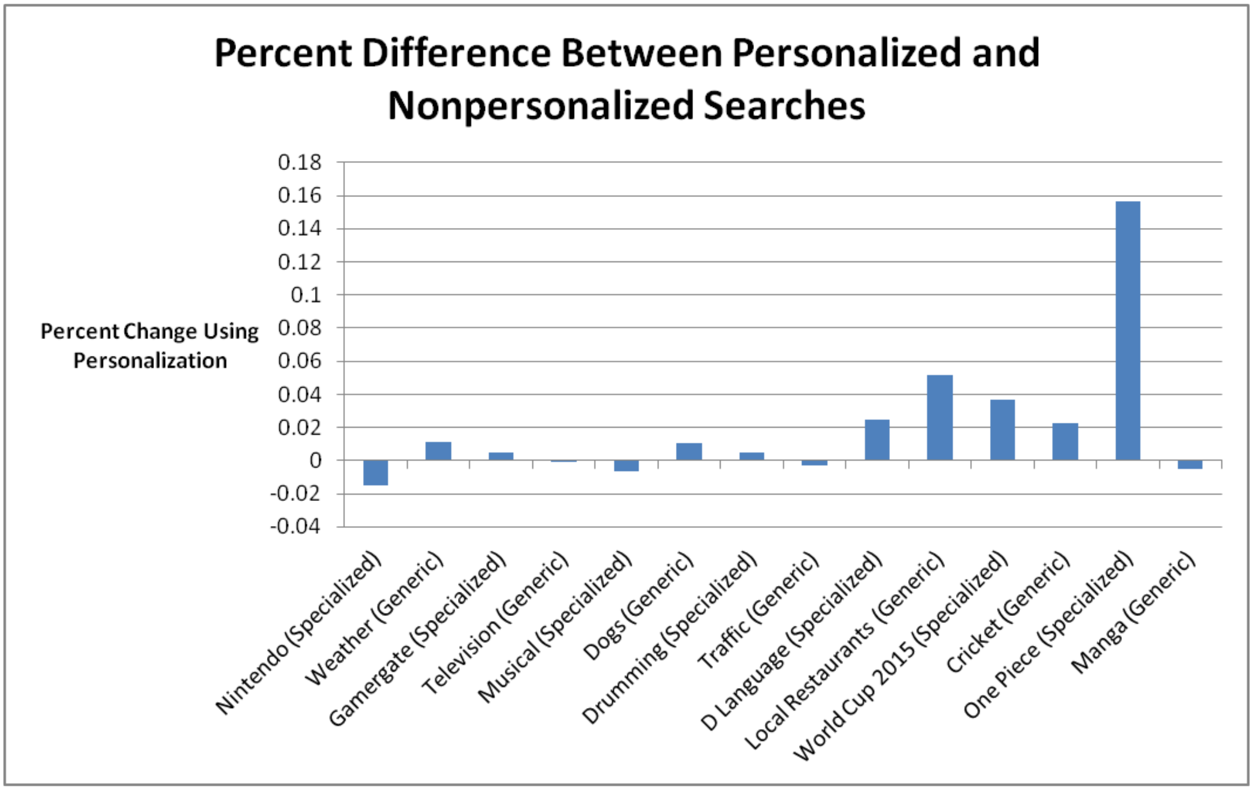
The final challenge was simply getting volunteers to run our bot. We had initially thought this would be easy since it wouldn't take too much time to record the url relevance values, but it ended up taking about 10 minutes to both run the bot and give the url relevance scores (which was enough to make it hard for us to find volunteers). While we believe we got enough data to make some conclusions, in the future we would likely need to consider compensating volunteers for their time in order to get more people involved in the study.

Results

We ran a total of 14 different searches (two per search user) with our bot between 2/21 and 4/10, recording both the personalized and standard results. Depending on the interval between searches and when the search started, this means each search has between 10 and 40 searches recorded (and total relevance scores). We took this data, ran some calculations on the results and graphed them to see the effects of personalization. Each graph is explained and discussed below. We computed two tailed paired t-tests with α=.05 between the total relevance scores and the percent changes to determine if there was a significant difference between the personalized and nonpersonalized results and their rates of change (specific t and n values included with the search's change over time graphs below).

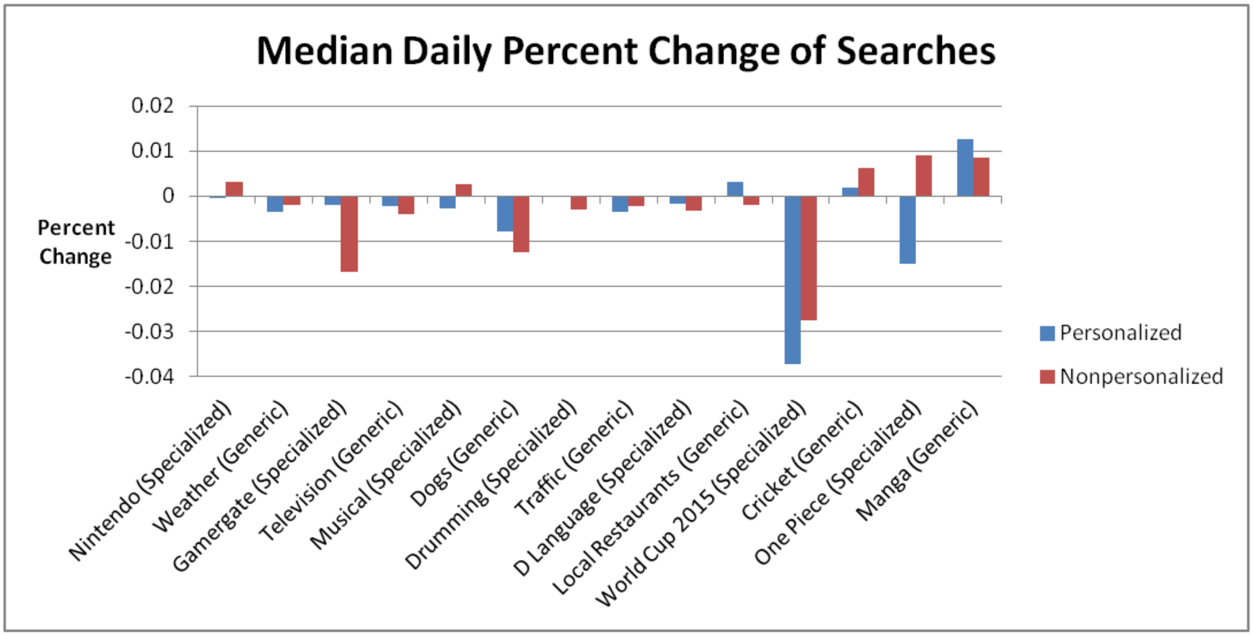


This graphs shows the median total relevance scores of the various personalized and nonpersoanlized searches we ran. From this graph, it appears that none of the searches had a significant difference (with the exception of the One Piece search). When we ran a paired t-test (details below) between the personalized and nonpersonalized searches however, we found that the Musical nonpersonalized search was significantly larger than the personalized search, and that the Dogs, D Language, World Cup 2015, Cricket and One Piece personalized searches were significantly larger than their nonpersonalized counterparts. We believe that the reason most of the searches had no major difference is because the first page of results (which has the highest weighted and most relevant web pages) is fairly static, so the largest portion of our total relevance score is the same between the two versions of the search. In addition, while there is some evidence of personalization on the later pages, it seemed to mainly be related to the changing positions of news stories, which tended to recieve a URL relevance score of 1 (since they are totally irrelevant) and not largely factor into our total relevance score. It is also worth noticing that personalization had an effect on 4 of 7 specialized searches, but only 2 of 7 generic searches. We believe this is because the specialized searches tend to be searched more often, so Google has more time to learn about them.

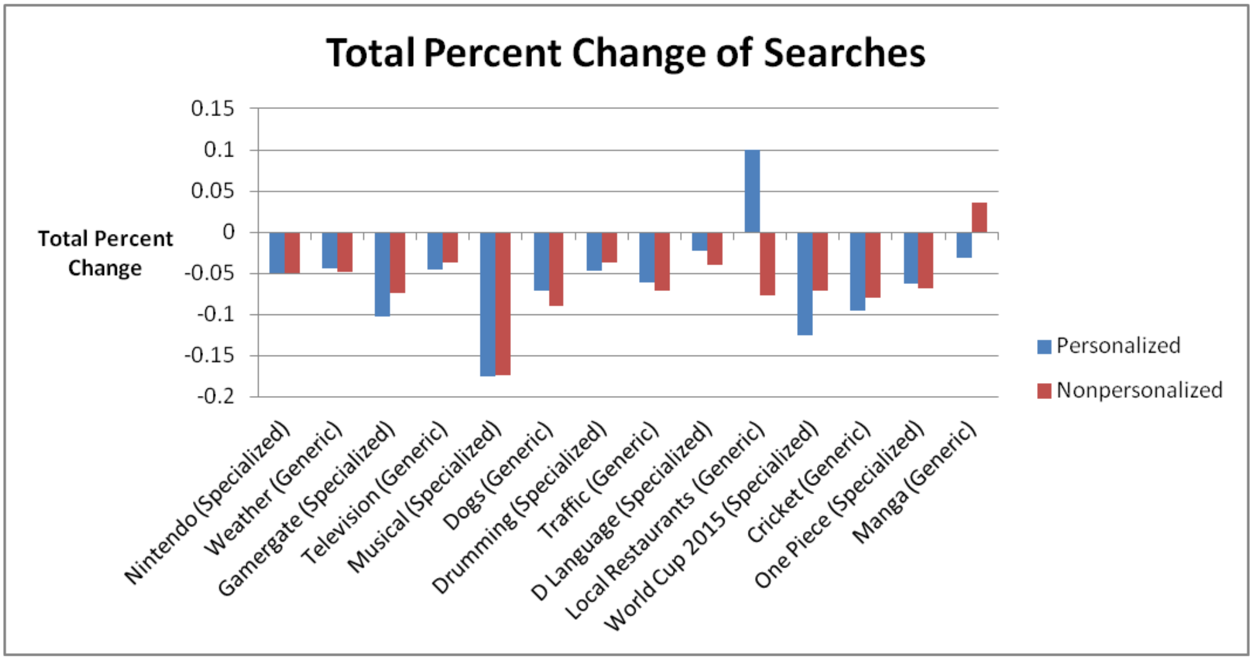


This graph shows the percent change between the median total relevance scores of personalized and non-personalized searches. Here, a positive change means personalization made the results more relevant, whereas a negative change means they made them less relevant. This further illustrates that most of our searches had an extremely small difference between the two, with most having a difference of less than 5%. The biggest exception is the One Piece search, where personalization significantly increased the relevance of the results results (with about a 15% increase in total relevance). The primary reason for this surge is that the user running this search had been following One Piece well over eight years and was logged into his Google account most of the time (for more details, see the One Piece change over time graph).

It is also interesting to note that our searches that had more data (closer to 40) had a much smaller change than those with less data (closer to 10). All the searches with about 40 data points had a total percent change of less than 2 percent, whereas most of those with around 10 points were higher than that. We believe that this may be because the the searches with more data had more time for the two versions of the search (personalized and nonpersonalized) to catch up with each other, allowing the median total relevance to stabilize close to each other, whereas there was not necessarily enough data for this to happen in the searches with fewer data points.



This graph shows the median daily percent change of the total relevance score for our various searches. From this graph, there doesn't seem to be any real difference between the rate personalized searches change and the rate nonpersonalized searches change (with the largest gap between the percent change of personalized and nonpersonalized being about 1.5%). This was confirmed with a paired t-test on the pairs of searches, where none of the results were significant. The only real interesting information from this graph is seeing which searches were most unstable (with total relevance rapidly changing over time) as they had a larger median percent change. For example, here we can see that the World Cup 2015 search had many steep changes since the median percent change was close to 3% downward.

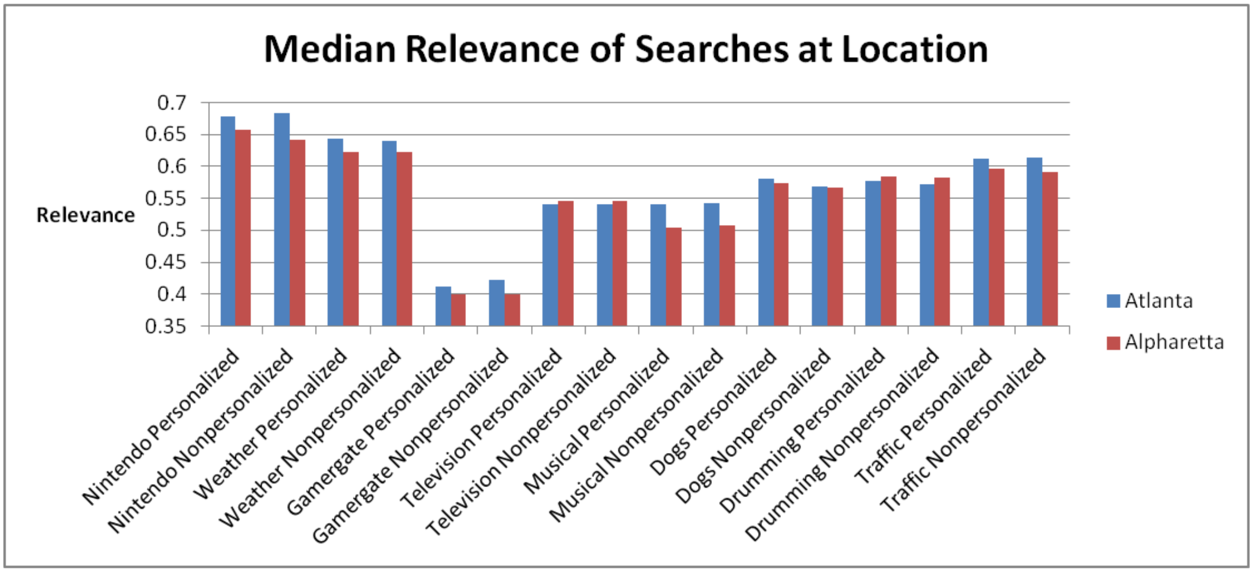


This is the total percent change in total relevance between the first search ran and the final search ran for each of the searches. From this graph we can see that there is no major difference in percent change between the personalized and nonpersonalized results for most of the results with the only exceptions being Local Restaurants and

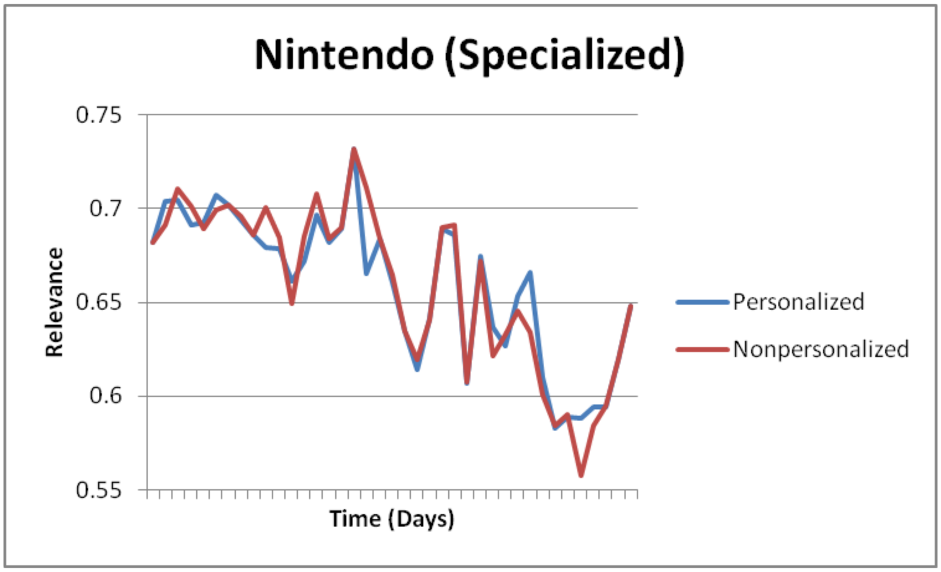
Manga. However, within this slight difference, it is interesting to note that the nonpersonalized results tended become less relevance more slowly than the personalized results for the specialized searches, (while there was no clear difference for the generic searches). This does make sense, as personalization would cause the search results of these commonly searched terms to change more quickly while Google tried to learn than in the nonpersonalized results.

We can also see that over time that the total relevance of almost all of the searches tended downwards. From examining the results, we found that this seemed to be caused by news stories. These stories tend to quickly pop up in the search results for a few days, before falling back off. They also tend to be completely irrelevant (the are almost always given a URL relevance score of 1), so they can really hurt the total relevance while they are in the results. The total downward trend is likely caused by the day we started the searches being a slow news day (so there were not many news stories in the results) and the final day having lots of news stories dragging the score down.

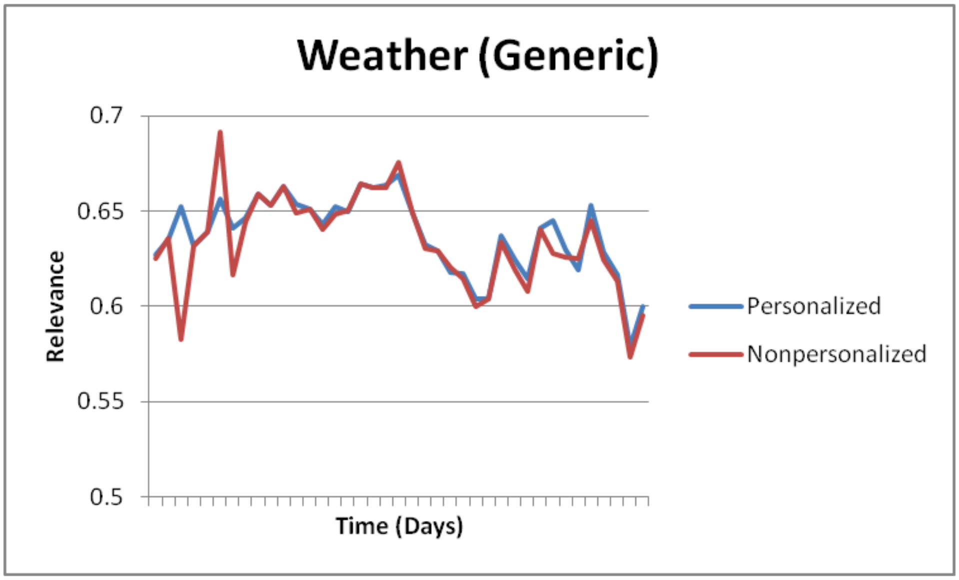
The biggest exception to the small change trend is the Local Restaurants search, which spiked in personalized search and sharply fell in nonpersonalized search on the final day. This seemed to be caused by a combination of one site with URL relevance 5 being missing from the original personalized search (http://tasteofatlanta.com/restaurants.php), whereas that site was in the personalized result and not the nonpersonalized result on the last day. This illustrates how effective one site the user finds relevant on the first page can be- enough to cause major changes in the total relevance score.



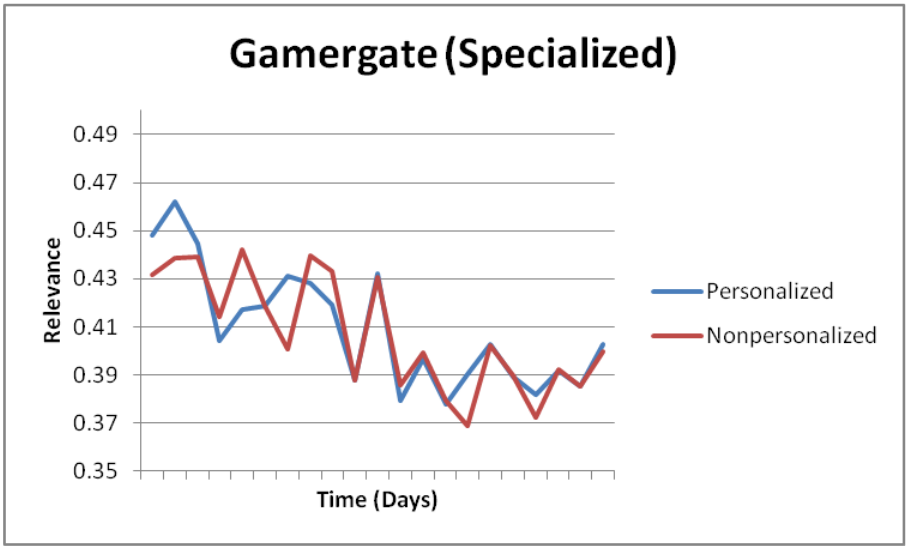
This graph shows the median total relevance of the searches we tracked the location of when searched in Atlanta and Alpharetta. We started tracking the location after our first presentation, where it was pointed out to us that Google also does some personalization based on where the search was performed. From our results, we could see no significant difference between the two results, with the largest difference being about .05 total relevance (which is small enough that the difference coming from the date of the search rather than the results themselves). This was confirmed using a two tailed two sample t-test (α=.05) on each pair of locations for a search. In the tests, the largest t score was around 2, but in that case it needed to be about 2.2 to be significant. The reason for this seems to fall into two categories. First, for most of these searches the location is completely irrelevant. For example, the results for the search “Nintendo” should not be changing between the two cities, as the location does not relate to the information the user is looking for. Second, for the few searches location does matter (like weather and traffic), Google tends to swap out one relevant result for an equally relevant result. For example, searching for weather in Atlanta will give you the Atlanta weather (URL relevance 5), while searching for weather in Alpharetta will instead give you Alpharetta weather (also URL relevance 5). Since the URL relevance score doesn’t change, neither does the total relevance, just the specific URL.



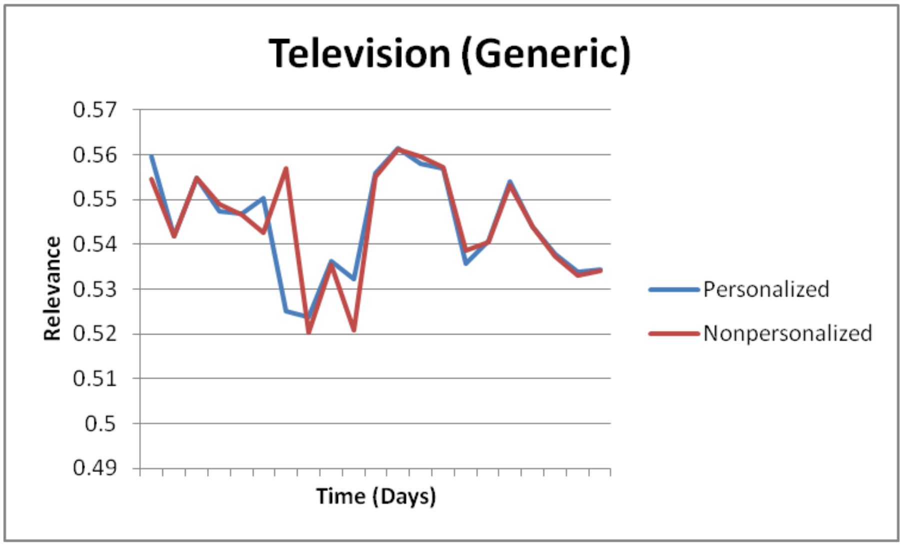
This graph shows the total relevance of the personalized and nonpersonalized Nintendo searches over time between 2/21 and 4/5 (relevance t-test t=.02, n=39, change t-test t=.003, n=39). Like most of our searches, there didn't seem to be any major difference between the personalized and nonpersonalized results. The most interesting part of this graph to note is the general downward trend of the results starting about halfway through the recorded period. From examining the results, we noticed that starting in mid March and continuing through early April, Nintendo made several major announcements that most of the internet thought was a really big deal, but the person running this search felt was irrelevant. This is a good example of why personalization could be so helpful- ideally, Google would have known that this person was not interested in news stories and not shown these results despite what the rest of the internet thought.



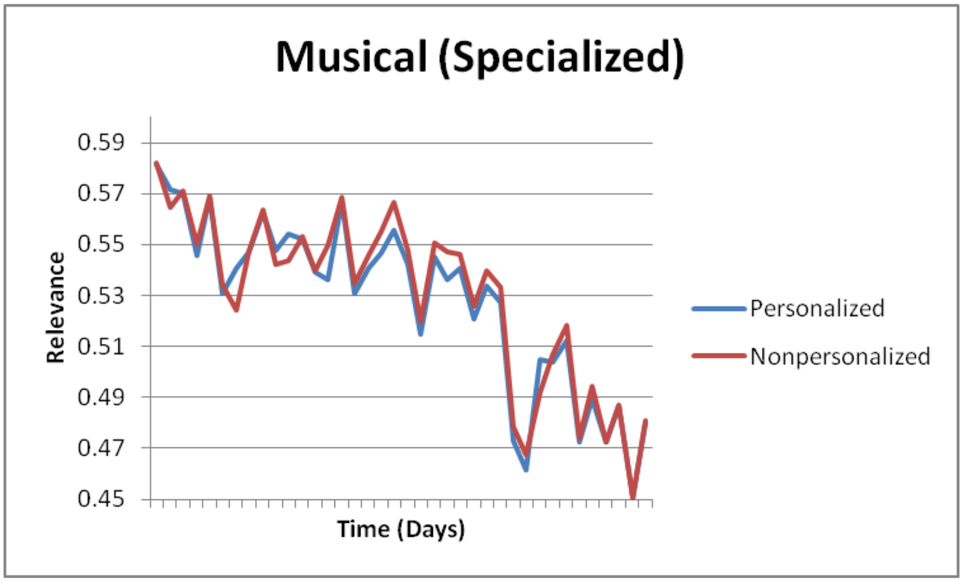
This graph shows the total relevance of the personalized and nonpersonalized Weather searches over time between 2/21 and 4/5 (relevance t-test t=1.54, n=39, change t-test t=-.055, n=38). While there is not a major difference between the personalized and nonpersonalized results of this search, it is interesting to note that the personalized results were consistently higher throughout (although only barely) and more stable near the beginning than the nonpersonalized results. From examining the URLs, we believe that the personalized weather results were higher and more stable because they had fewer news stories show up in the results, so the top five pages of results did not change much and were not dragged down by irrelevant news.



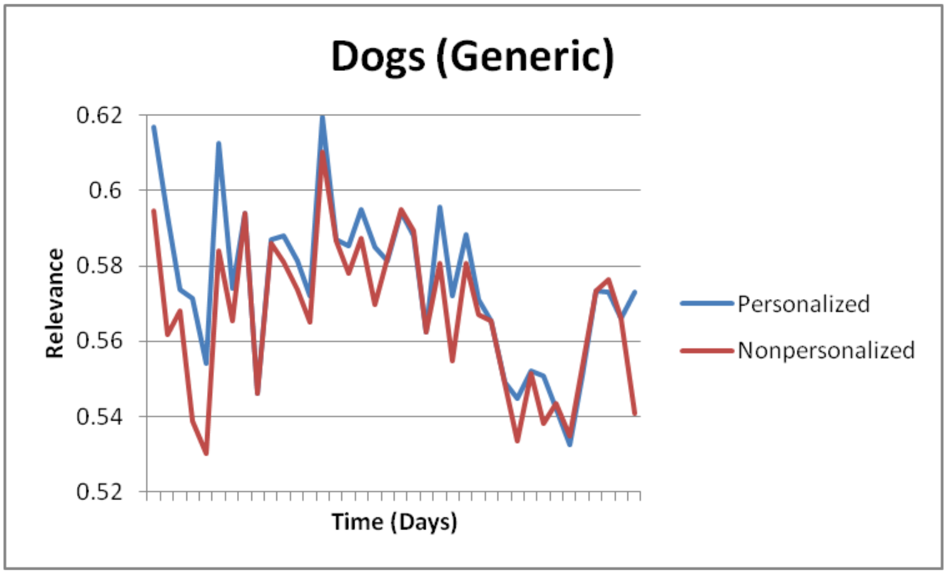
This graph shows the total relevance of the personalized and nonpersonalized Gamergate searches over time between 3/15 and 4/5 (relevance t-test t=.66, n=21, change t-test t=-.210, n=20). We felt this would be an interesting search topic because there are two extremely different meanings you can find for it online, so it was ripe for personalization. Despite this, we once again found no real difference between the personalized and nonpersonalized results. In addition, these results easily had the lowest total relevance scores of any search. From examining the results, both of these problems had the same cause- the vast majority of the results were from news sites that were using the wrong meaning of the term for the user running the search. Ideally Google should have altered the results for personalization, but instead it simply gave the one sided results because they were from more popular (and thus, more widely linked) sites.



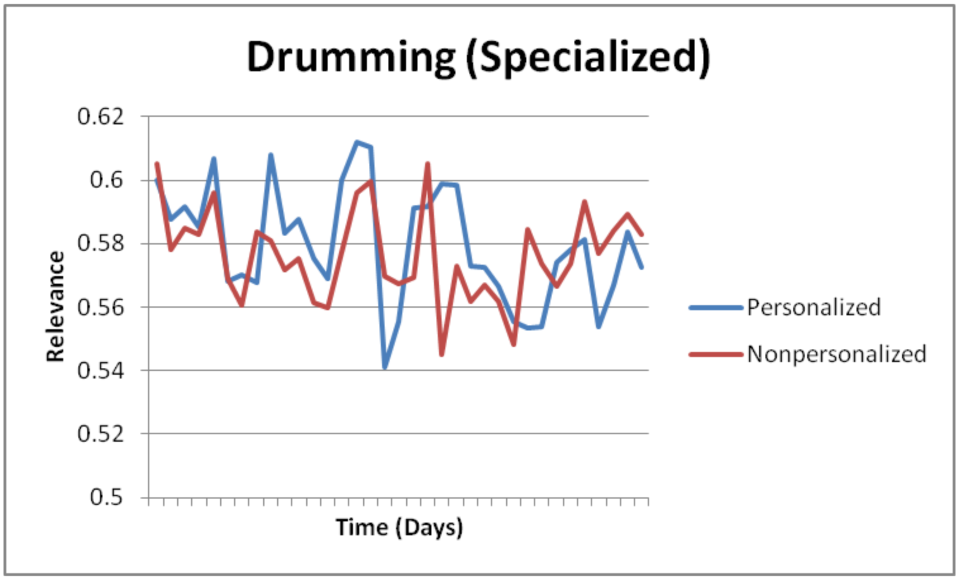
This graph shows the total relevance of the personalized and nonpersonalized Television searches over time between 3/15 and 4/5 (relevance t-test t=-.15, n=21, change t-test t=-.105, n=20). Once again, this graph shows no major difference between the two types of searches. What is most interesting here is how little the results varied over time in total relevance. Despite not being relatively static like the weather results were, the television total relevance scores didn't vary by more than about .04 (one of the smallest differences we encountered). From the results, it appears that this was caused by the new results being equally irrelevant with those they were replacing (for example, one television manufacturer's website being replace with a different manufacturer's site).



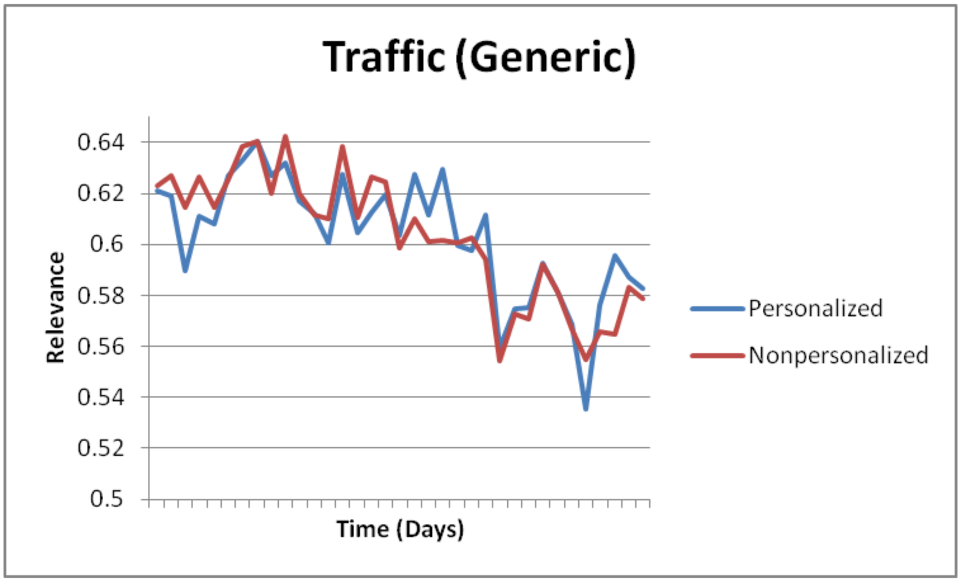
This graph shows the total relevance of the personalized and nonpersonalized Musical searches over time between 2/21 and 4/5 (relevance t-test t=-2.16, n=38, change t-test t=-.005, n=37). Here, the nonpersonalized search results were significantly better than the personalized results. The other interesting part is that this topic had the sharpest dive in total relevance between the first day and the last day. This dive seemed to be caused by a large increase in websites for obscure musicals showing up in the results, while the early search results were largely sites about musicals in general (which this user found much more relevant). This also appeared to be what made the nonpersonalized results better than the standard results, because Google returned the obscure musicals higher on the user's personalized search results.



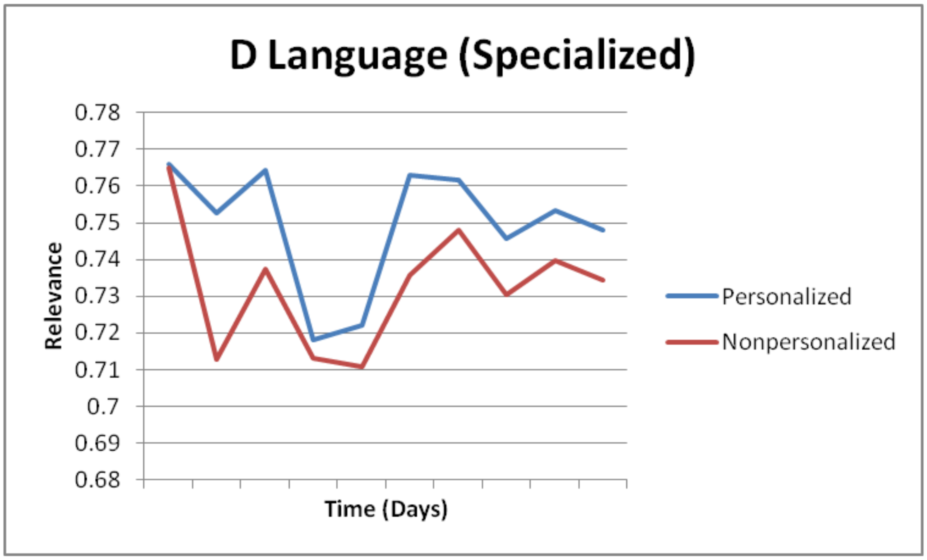
This graph shows the total relevance of the personalized and nonpersonalized Dogs searches over time between 2/21 and 4/5 (relevance t-test t=4.66, n=38, change t-test t=.165, n=37). Here personalization significantly improved the relevance of the search results. The person who ran this search considered sites that made him think of cute dogs most relevant and news stories about dogs least relevant, and Google seemed to pick up on that enough to have the personalized total relevance score remain above the nonpersonalized relevance for most of this graph. It is also worth noting that this difference is not as significant as we reported in our April 1 presentation, as we appear to have made a math error on those results that was causing the larger difference.



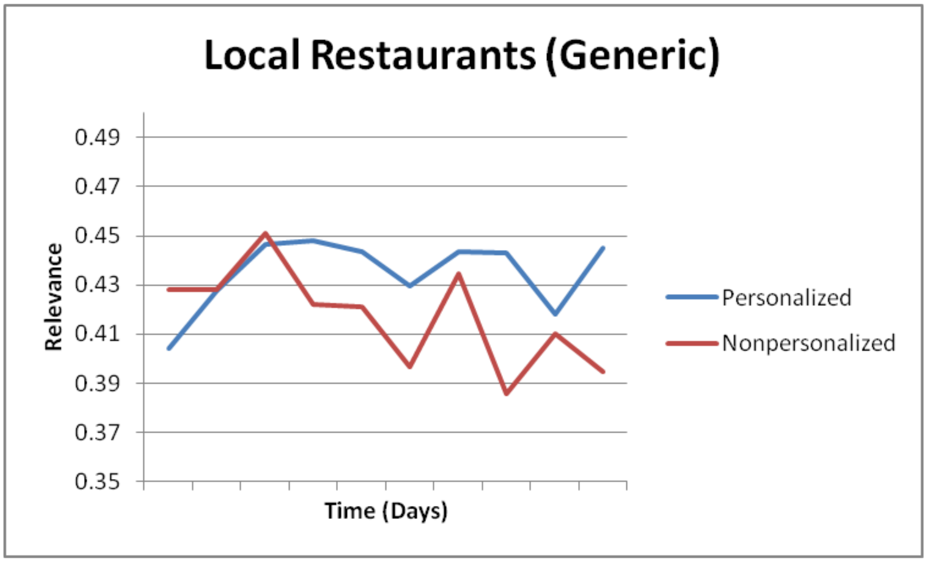
This graph shows the total relevance of the personalized and nonpersonalized Drumming searches over time between 2/21 and 4/5 (relevance t-test t=1.04, n=35, change t-test t=-.032, n=34). This was another instance where personalization had a slightly positive benefits over the nonpersonalized results (with the personalized results beating the nonpersonalized results by about .02 for a decent amount of the recorded searches), although not enough so to be significant. The slight difference in this case seems to be that the personalized search had a few more instances of useful drumming tips, whereas the standard results had a few more advertisements for drumming lessons and drum related news.



This graph shows the total relevance of the personalized and nonpersonalized Traffic searches over time between 2/21 and 4/5 (relevance t-test t=.05, n=37, change t-test t=.135, n=36). Here, there is not a significant difference between the personalized and nonpersonalized relevance. From looking at the results, the small difference seems to be caused by the person running the search deciding that the cities mentioned in one search were slightly better than the cities in the other search for no clear reason. This means they would have caused the difference in the score but would not have helped Google better personalize the results over time.



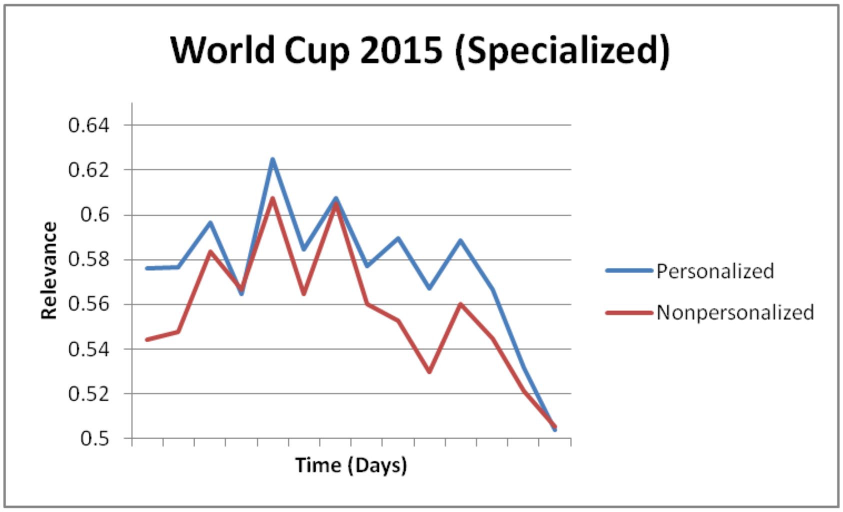
This graph shows the total relevance of the personalized and nonpersonalized D Language searches over time between 3/11 and 4/10 (relevance t-test t=4.63, n=10, change t-test t=.232, n=9). This instance shows a significant benefit to personalization, with the personalized search consistently being ahead of the nonpersonalized results. In addition, this had the highest total relevance score of any of the searches. Unlike many of the other searches, the user running the search saw pages that they had visited frequently before in the results of the personalized search that did not show up as much in the nonpersonalized results. This lead to the benefit of personalization and high total relevance.



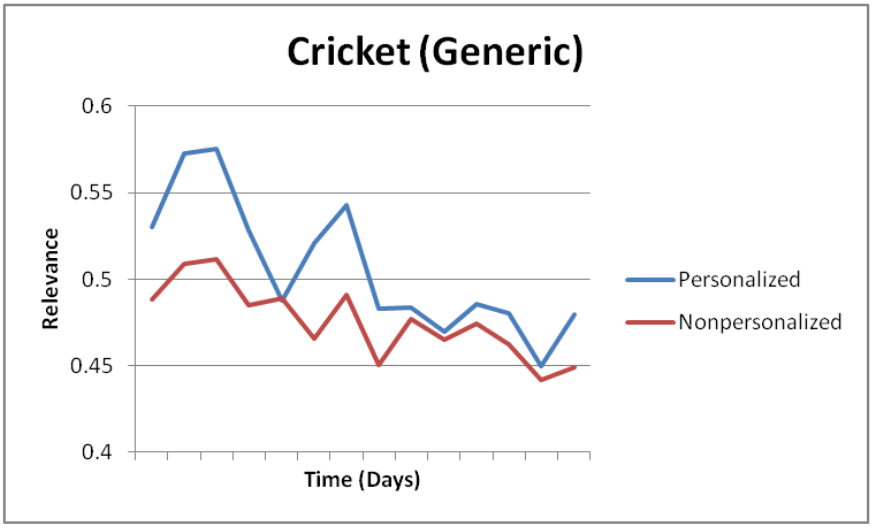
This graph shows the total relevance of the personalized and nonpersonalized Local Restaurants searches over time between 3/11 and 4/10 (relevance t-test t=2.25, n=10, change t-test t=.714, n=9). Here we can see that the personalized results were consistently more relevant than the nonpersonalized results, but not by a significant amount. In addition, the overall relevance was extremely low because after the first page of results, none of the results seemed to location aware and were not actually local to the user. It seems that this is a search that could have greatly benefited from more location based personalization, like that seen in Weather and Traffic.

There was a paradigm shift in the approach used to run tests and gather the CSVs on Cricket, World Cup 2015, Manga and One Piece searches. We ran the bot on a computer where Mozilla Firefox was not the default browser with the intention of adopting a clean slate approach. We did not want the previous search history in the browser (cookies and the web browser cache) to skew the results of the searches since we wanted both personalised and nonpersonalised search to start from the same point, independent of the previous searches on the browser.

Another important feature of this last set of searches was that the specialized and generic searches were both on closely related topics unlike the previous searches. For instance World Cup 2015 refers to the recently concluded cricket world cup hosted by Australia and New Zealand. Hence, World Cup 2015 and Cricket were closely related. Again One Piece is a Japanese manga series serialised in the Weekly Shonen Jump magazine. Hence, One Piece and Manga were closely related search topics.

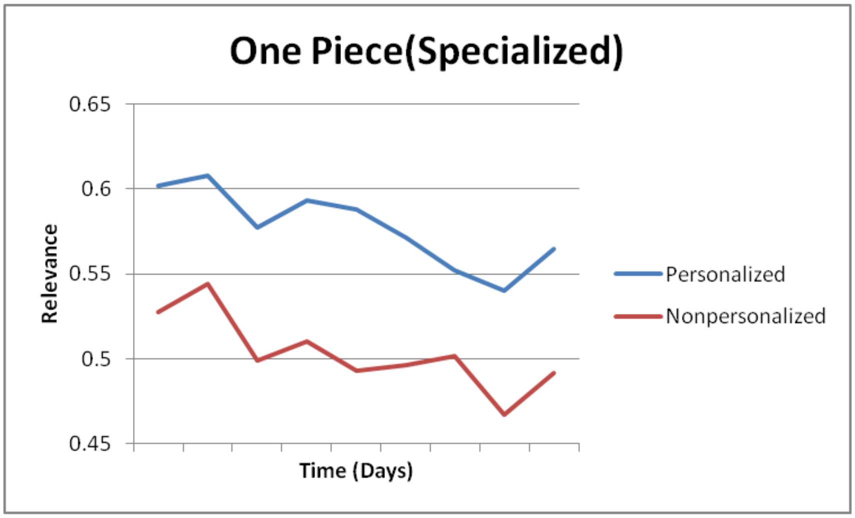


This graph shows the total relevance of the personalized and nonpersonalized World Cup 2015 searches over time between 3/22 and 4/9 (relevance t-test t=5.31, n=14, change t-test t=-.703, n=13). Here the personalized results were significantly better than the nonpersonalized results. One important feature of this plot is the periodic crests and troughs. To understand the reason for this behavior, we must take into account the following points. Firstly, the World Cup 2015 was the cricket world cup which took place earlier this year. Secondly, the person who ran this search was supporting India in the world cup and the matches involving the Indian cricket team were scheduled at periodic intervals. The person running these searches would use Google to find out the score and also live stream the match at times. Hence, he ran multiple searches on the days when India actually played and those were the days when he observed a jump in improvement. Another interesting trend was the sudden drop in relevance towards the end. This was probably due to the fact that his interest waned once India was eliminated from the world cup.

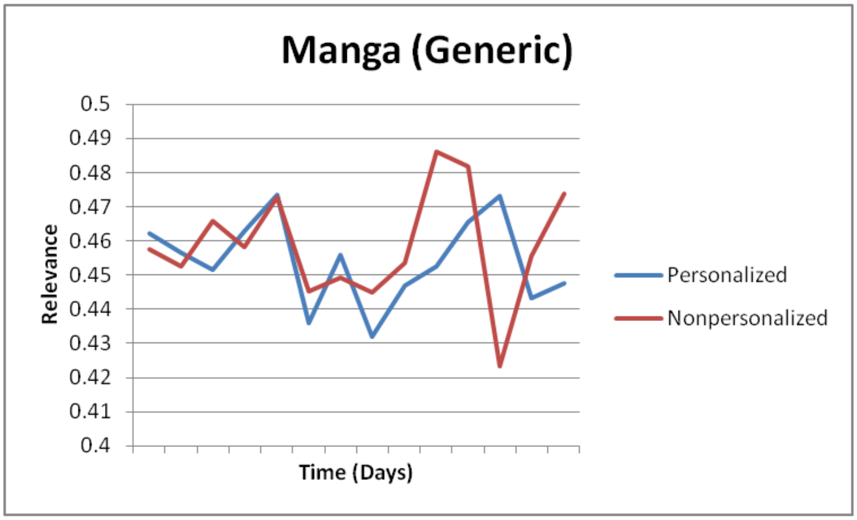


This graph shows the total relevance of the personalized and nonpersonalized Cricket searches over time between 3/22 and 4/9 (relevance t-test t=4.98, n=14, change t-test t=-.030, n=13). In this plot we saw that personalised search results consistently performed significantly better than nonpersonalised search over the entire time span. The was due to the inter-relationship between the search on World Cup 2015 and cricket. When this search was run, the user primarily searched for World Cup updates and these came under the broader category of cricket searches. In personalised search, many of the results were live streaming links and online commentary pages related to the ongoing

World Cup whereas in non-personalised search there were numerous generic Wikipedia pages on the origin and rules of cricket in addition to the live streaming links thereby, decreasing the overall relevance on any particular day.



This graph shows the total relevance of the personalized and nonpersonalized One Piece searches over time between 3/22 and 4/1 (relevance t-test t=17.94, n=9, change t-test t=.015, n=8). Here we can see that personalisation provided a significant improvement in relevance over non-personalised search, with about a 15% increase in relevance from personalization. This was because the user who ran this search had been reading the manga One Piece and watching the anime over a period of around nine years. Secondly, Google Chrome was his default browser on which he was always logged into his Google account. Hence, Google personalised search had a plethora of results at its disposal (personalised data linked to the Google account) thereby greatly improving the relevance of personalised search. On the other hand as mentioned earlier this set of searches was performed on Firefox using a clean slate approach and nonpersonalised search did not have any browser data prior to the date of starting the search to fall back upon.



This graph shows the total relevance of the personalized and non-personalized Manga searches over time between 3/22 and 4/9 (relevance t-test t=-.83, n=14, change t-test t=-.316, n=13). In this plot the relevance values of both personalised and non-personalised search are pretty close most of the time, giving no significant difference between the two. One interesting feature in this plot that immediately caught our attention was the sudden rise in relevance on the last day of running this search. This day happened to be a Thursday and the Weekly Shonen Jump

(Japanese manga) releases a new episode on every Thursday. As mentioned earlier the person running this set of searches was an avid follower of manga (One Piece in particular) and hence observed numerous relevant results on that day as compared to the previous few days. That is probably one of the reasons for the periodic peaks in the plot over a longer time span. Lastly, to take into the account the surge in relevance on Thursdays, the manga search was performed over a longer period of time as compared to the previous searches in this set.

Future Work

There are a few interesting ways this experiment could be expanded. First, and most obviously, we could collect more data over a longer period of time. For our project. we collected data over the course of about 1.5 months, which is long enough to start to see some trends but not long enough to study anything extremely long term. If we were able to run the same tests over the course of an entire year, we may get some more interesting and better distinguished results.

Second, we could refine the searches to try and avoid some of the problems we ran into with this project. Here, we let the person doing the searches pick the terms so that they could cater to their interests, but some of the searches they picked were often broad enough for there to be news stories and specific enough that the first page was hard to personalize. On the other hand, searches like World Cup 2015 and One Piece were specific enough that we could start to see a difference between the two results. Spending more time selecting better searches would likely give us better results overall.

Finally, we could run this bot on a different search engine. When we were planning our project, we discovered that Bing seems to do a lot more work on personalization than Google does [2]. We used Google for this project out of necessity (since no one we know uses Bing), but we would have likely gotten more interesting results by using that site instead. In the future, we could have participants use Bing just for the duration of the study and see if we get a bigger difference from that search engine.

Conclusion

From the data we gathered for the past 1.5 months, we saw some significant differences between personalized and nonpersonalized search results, but most of the searches had no major differences between the two. This largely appears to be due to the most relevant information already being on the first page of the search results regardless of personalization. While personalization has a larger effect on the later pages in a search, since most people will only look at the first page before refining their search, it makes sense that if the first page was similar the two searches would have similar scores. We also discovered that personalization has a larger effect on specialized search results (with 4 of 7 results being significant) than generic searches (with only 2 of 7 being significant). This makes sense, as the specialized searches are used often enough that Google can learn more about the users than with the generic searches.

In addition, there seemed to be no major difference between the rates at which the personalized and nonpersonalized results change, with none of the rates having a significant difference in the rates of change over time. For any given topic, the specific rate of change seemed to have more to do with how many news stories could be related to the topic than anything else. If there were many news stories, the results would fluctuate a lot more over time, whereas searches with few news stories would remain relatively static. Since the same news showed up in both types of searches, their rates of change were also the same.

References

[1] Browser Automation. (n.d.). Retrieved February 2, 2015, from http://www.seleniumhq.org/

[2] Crook, A., & Ahari, S. (2011, February 10). Making search yours. Retrieved April 17, 2015, from http://blogs.bing.com/search/2011/02/10/making-search-yours/

[3] Dou, Z., Song, R., & Wen, J. R. (2007, May). A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web* (pp. 581-590).

[4] Pretschner, A., & Gauch, S. (1999). Ontology based personalized search. In*Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on* (pp. 391-398). IEEE.

[5] Search Engine Optimization Starter Guide. (2010, January 1). Retrieved February 2, 2015, from http://static.googleusercontent.com/media/www.google.com/en/us/webmasters/docs/search-engine-optimization-starter-guide.pdf

[6] Speretta, M., & Gauch, S. (2005, September). Personalized search based on user search histories. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on* (pp. 622-628). IEEE.