Demographic Information Flows

Ingmar Weber
Yahoo! Research Barcelona
Av. Diagonal 177
08018 Barcelona, Spain
ingmar@yahoo-inc.com

Alejandro Jaimes
Yahoo! Research Barcelona
Av. Diagonal 177
08018 Barcelona, Spain
ajaimes@yahoo-inc.com

ABSTRACT

In advertising and content relevancy prediction it is important to understand whether, over time, information that reaches one demographic group spreads to others. In this paper we analyze the query log of a large U.S. web search engine to determine whether the same queries are performed by different demographic groups at different times, particularly when there are query bursts. We obtain aggregate demographic features from user-provided registration information (gender, birth year, ZIP code), U.S. census data, and election results. Given certain queries, we examine trends (from high to low and vice versa) and changes in the statistical *spread* of the demographic features of users that issue the queries over time periods that include query bursts. Our analysis shows that for certain types of queries (movies and news) distinct demographic groups perform searches at different times, suggesting that information related to such queries *flows* between them. Queries of movie titles, for instance, tend to be issued first by young and then by older users, where a sudden jump in age occurs upon the movie's release. To the best of our knowledge, this is the first time this problem has been studied using search query logs.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Experimentation, Human Factors, Measurement

Keywords

demographics, trends, query logs

1. INTRODUCTION

Information diffusion has been a topic of interest in sociology, marketing, consumer behavior, management, and the health sciences, among several other fields since at least the 1960s [9]. Many dimensions have been studied, including the application of diffusion theories to consumer research, for targeting of new products and services, and for modeling the diffusion of innovation in social systems. Understanding information flows across demographic groups is particularly important in areas such as public health policy and can be useful for studying information inequality. There are applications in advertising [3] and other fields.

In this paper, we analyze the query logs of a large U.S. search engine to study information flows across demographic groups. We define information flows as changes in the aggregate distribution of demographic features of groups of registered users that perform given queries within a time window in which there is a query burst. In particular, we consider changes in the mean of demographic features (e.g., from young to old) as well as in the spread (e.g., from a niche to the whole population) for movies and U.S. news. We obtain demographic information by linking user provided registration information (gender, birth year, ZIP code), U.S. census data, and election results to obtain per-ZIP-code estimates of variables such as income and race.

Our contribution is thus to examine information flows by analyzing a large-scale query log in terms of the demographic characteristics of the groups searching. The results of our study suggest that queries "travel"[1] between different demographic groups in the categories of movies and news.

2. RELATED WORK

Recent work on information diffusion and propagation has been carried out on the web in the context of news [8], blogs [1, 6], and social networks. Leskovec et al. [8], for instance, track short distinctive phrases through on-line text. They automatically cluster variants of such phrases and examine the resulting news cycle (i.e., they analyze temporal patterns in the spread of those phrases in blog posts and news articles). Gruhl et al. [6] study topic propagation in blogs, making a distinction between horizon models (i.e., those that change over the course of months, years or decades) and snapshot models (i.e., those that change over days, weeks or months). They characterize spikes in the data (i.e., by examining aggregates of posts at the topic level) and model the path of topics through individuals in the blogspace (i.e., by classifying individuals in terms of how they contribute to the propagation of information). Adar et al. [1] study informa-

[1]Strictly speaking, we do not know *why* queries might be issued by one demographic group and then another one, thus the words "travel" and "flow" must be interpreted with caution as we have no evidence of flow of information *between* such groups.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

tion propagation by determining the path that information takes in blog networks.

There has been a considerable amount of work on analyzing web search queries (without considering demographics). This includes examining query volume bursts over time [11, 4], and changes in the distribution of topics covered by the queries at several levels (e.g., Passetal et al. [10] analyze variations within a day, and Beitzel et al. perform hourly analysis [2]).

Finally, Jones et al. infer demographics from query logs [7] and Cheng et al. focus on ad click prediction [3]. Weber et al. [13] studied the static demographic distribution of a large U.S. search engine, but did not consider changes over time or in volume.

Conceptually, our work can be seen as a generalization of Google Flu Trends [5], but rather than tracking the change in distribution across different geographical regions, we track changes along demographic dimensions.

To the best of our knowledge this is the first time that information *flows* between different demographic groups, i.e. *changes* in the distribution as a function of time or query volume, are studied using search engine query logs.

3. DATA SET & METHODOLOGY

The query log we extracted is a sample of web search queries by registered users for a large U.S. web search engine for the period from Mid-April 2009 to Mid-May 2010. The total sample, after filtering out non-registered users or users who provided a non-existing ZIP code, contains several hundred million queries. No personal information was used and all of the analysis was done in aggregate. For our analysis we used a small subset of these queries, namely only those 20 distinct queries listed in Table 2.

The queries used for our study were selected such that (i) they were not in the extreme "head" of the distribution (such as `facebook` or `youtube`) but were still frequent enough to allow statistically significant observations, (ii) one could expect a certain time-dependency in the query volume and demographic distribution during the investigated time period, and (iii) queries could be assumed to be unambiguous (e.g. we dropped `avatar` from an earlier selection as it often seemed to refer to "yahoo avatar" and not the movie).

The query log data contains for each query a timestamp T and demographic information, namely, the birth year, the gender and the ZIP code, obtained from user-provided registration. The ZIP code is used to obtain additional information about the users from the 2000 U.S. census[2]. Each query is therefore annotated with a timestamp and estimated expected values for each of the twelve features shown in Table 1.

We want to detect if for a given query there is correlation/trend between the timestamp and one or several of the feature values. Our goal is not to detect bursts or periodic patterns [11], but rather to detect trends (from high to low or vice versa) in demographic features as well as changes in the statistical *spread* of the corresponding feature.

As automated queries were removed prior to our analysis, we decided to *keep* queries repeated by the same user, even though such repeat queries account for about 40% of our query volume. We preferred this option as our whole analysis is done on a per-query basis, as opposed to a per-user basis, and as advertising is also sold on a per-query basis.

[2]`http://factfinder.census.gov/`

Feature	Abbr.	Average	U.S. average
av. household size	hhs	2.63	2.59
mean travel time (m)	tt	25.8	25.5
non-engl. language (%)	ne	16.7	17.9
BA degree (%)	ba	26.8	24.4
Below poverty (%)	bp	10.5	12.4
Per-capita income ($)	inc	23,578	21,587
White (%)	wh	78.6	75.1
Black (%)	bl	9.6	12.3
Asian (%)	as	4.0	3.6
Hispanic (%)	his	10.9	12.5
birth year	by	1967*	1975*
gender (% male)	g	52.0	49.1

Table 1: List of the 12 demographic features we used in our study with both our per-query averages and the U.S. census per-person averages. The median is reported for the birth year.

4. TRENDS OVER TIME

We analyze how (i) the demographic middle and (ii) the demographic spread of queries change over time. "Demographic middle" refers to the average feature value F at a point in time T. For example, the average birth year of a query such as `avatar movie` changes from a high birth year (young) to low (old). By "demographic spread" we mean the standard error $\sigma(F)$ of the feature values at a point in time T. For example, the query `crazy heart` sees a significant increase in its spread for most features as time goes on.

We use linear regression to detect changes both in (i) the average and (ii) the spread of the feature values F over time. In the first case, it is a direct application of the text book algorithm [12]. In the second case, we first need to group several queries and their (T, F) pairs into buckets of B days as we cannot estimate the spread $\sigma(F)$ for a single pair (T, F). The average value of T within a bucket then becomes the timestamp for the whole bucket and the average value of F within the bucket becomes the corresponding feature value. A smaller bucket size B corresponds to a higher time resolution and is more appropriate for fast changing topics such as movies. A large value of B is more appropriate to analyze slower "flows" for topics such as `health care reform`.

For each of the 20 queries and each of our 12 features we computed a least-squares linear regression for both (i) (T, F) and (ii) $(T, \sigma(F))$ pairs. We used models with a constant offset C of the form $Y = W \cdot T + C$, where Y is the variable to be predicted, namely either F or $\sigma(F)$. Apart from the actual size of a linear trend W, the model also gave us a t-score to quantify the model's confidence that $W \neq 0$. In all of our analysis, we only considered trends to be statistically significant if this t-score surpassed 2.58, corresponding to a confidence level of 1% in a two-sided hypothesis test.

For each query, the "T vs. F" column in Table 2 gives the two trends with the highest, significant t-score (if any). For example, for `avatar movie` the feature "by" (birth year, see abbreviations in Table 1) had a t-score of 32.1 and the fitted line had a slope of -18.9 (in terms of birth year) over one year. So, according to the linear fit, the average age increased by almost 19 years during 12 months, and for the News queries, birth year is just as likely to have a positive or negative trend (3/10 queries).

Concerning the detection of changes in the spread ($\sigma(F)$)

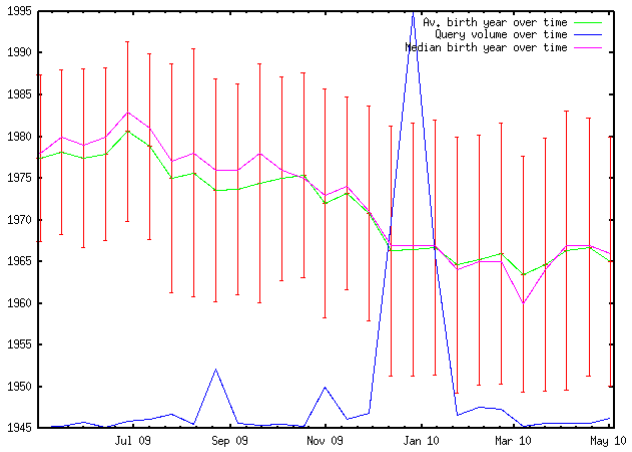

Figure 1: The query `avatar movie` exhibits a general trend from high birth year (young) to low (old). This trend starts before the burst caused by the release of the movie on Dec. 18, 2009, but the increase in volume seems to act as a catalyst and also causes a spread in terms of the error bars $\sigma(F)$. After the movie's release the distribution appears to be comparatively stable. In these figures the Y axis does not correspond to actual absolute query volume.

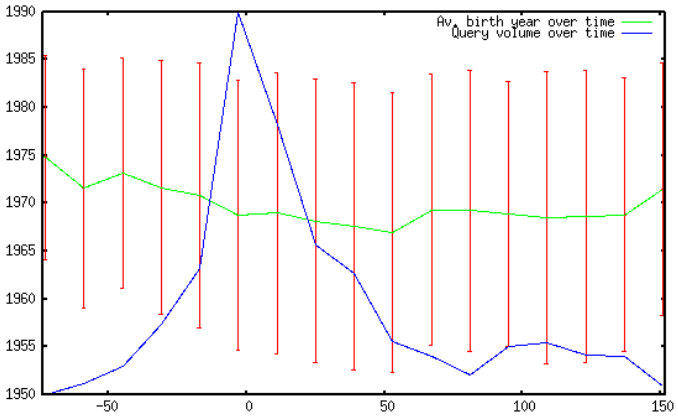

Figure 2: Plot of the birth year over time for the combination of all 10 movies indicating a common shift in birth year, from young to old. Movies were aligned according to their release date and the x-axis shows days before and after the release. The volume is the macro-average of the normalized volume. As for the individual query `avatar movie`, the size of the error bars increase coinciding with the release date.

of the distribution, column "T vs. $\sigma(F)$" in Table 2 gives the number of features with a positive trend, i.e. where the spread has increased over the course of a year, as well as the number of features with a negative trend. Figure 1 shows the behavior of the birth year over time for the query `avatar movie`. Apart from the aforementioned trend, one can also see an increase in the size of the error bars.

5. COMMON PRE-/POST-BURST BEHAVIOR

Many movies are only known within a niche of the population before they are released. When movies are released nationally, they are widely advertised, which causes a burst in interest which, along with query volume, spreads across a more diverse range of user segments. In order to detect commonalities across the flows for all ten movies, we aligned the time axes using the movies' release dates. After this shift in time, for each time bucket the value F of each feature was taken to be the (macro-) average across all 10 queries in the set and the same was done for V and $\sigma(F)$. Only buckets with at least 9 queries contributing were used. For each movie we computed its maximum query volume V in any bucket of 14 days. This volume was then used to normalize the volume in all buckets to the range $[0.0, 1.0]$.

Figure 2 shows a plot of birth year vs. time for this combined data for a range of 80 days before and 150 days after the release date. Though the trend is not as pronounced as for the individual case of `avatar movie`, it is still noticeable and coincides with the results of [4].

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented a study of information flows between different demographic groups by analyzing the query log of a large U.S. web search engine. Demographic information is linked to queries by using registration information provided by users (gender, birth year, ZIP code) and joining it with U.S. census information to obtain per-ZIP-code estimates of variables such as income and race distribution. Our analysis shows that whereas some queries (e.g., `guantanamo bay`) have "converged" with respect to their demographic distribution, the majority of queries considered shows statistically significant differences in at least one demographic feature over the the time period analyzed. Furthermore, the behavior for queries in our movies category tended to be similar with a consistent "flow" from younger to older users.

Although our initial data set included several hundred million queries, we based our analysis on a set of only 20 queries. Future work includes automating parts of the process to do a larger scale analysis, in particular clustering queries per topic, and a deeper analysis within categories (e.g., building prediction models for marketing or informing public policy to alleviate information inequality).

7. REFERENCES

[1] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence (WI)*, pages 207 – 214, 2005.

[2] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Conference on Research and development in information retrieval (SIGIR)*, pages 321–328, 2004.

[3] H. Cheng and E. Cantú-Paz. Personalized click prediction in sponsored search. In *Conference on Web search and data mining (WSDM)*, pages 351–360, 2010.

[4] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences (PNAS)*, 105(41):15649–15653, 2008.

[5] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza

Topic	Query	Volume	T vs. F	T vs. $\sigma(F)$	B	origin
Movies	avatar movie	+++	by, -18.9, 32.1; tt, -1.1, 4.3	2+ / 0-	14	12/18/09
	up	+++	ba, -2.9, 6.1; inc, -1461, 4.6	0+ / 0-		05/29/09
	the hurt locker	++	g, -23.1, 11.7; by, -6.2, 10.3	2+ / 0-		06/26/09
	crazy heart	++	g, -70.3, 14.1; ba, -17.1, 10.5	10+ / 0-		12/16/09
	ninja assassin	++	tt, -2.8, 5.0; wh, 8.6, 4.0	0+ / 0-		11/25/09
	planet 51	++	-	0+ / 0-		11/20/09
	nine movie	++	by, -21.4, 11.1; his, -5.4, 2.6	0+ / 0-		12/18/09
	a serious man	++	ba, -9.2, 4.2; g, -21.4, 3.7	5+ / 0-		10/02/09
	invictus movie	+	g, -60.4, 4.5; ba, -18.3, 4.1	6+ / 0-		12/11/09
	everybody's fine	+	inc, -6838, 2.8	3+ / 0-		12/04/09
U.S. News	sarah palin	++++	by, -1.1, 7.7; hhs, -.02, 6.2	1+ / 0-	28	-
	health care reform	+++	inc, 1471, 6.3; g, 6.5, 5.9	0+ / 0-		-
	barack obama	+++	by, -3.1, 13.2; inc, 1554, 10.4	3+ / 1-		-
	sonia sotomayor	+++	by, 3.9, 9.2; ba, -2.0, 4.5	1+ / 0-		-
	fort hood	+++	by, -12.1, 12.8; inc, 5381, 9.6	1+ / 0-		-
	al franken	++	his, -2.0, 3.5; ne, -2.2, 3.5	0+ / 3-		-
	joe wilson	++	-	0+ / 1-		-
	ben bernanke	++	by, 4.7, 5.1; hhs, -0.1, 2.5	0+ / 0-		-
	guantanamo bay	++	-	0+ / 4-		-
	same sex marriage	++	bp, 2.2, 3.1; ba, -3.0, 2.6	0+ / 0-		-

Table 2: Our query set, grouped into two topics. The volume indicators "+" are on a logarithmic scale. The results for columns "T vs. F" and "T vs. $\sigma(F)$" are discussed in Section 4. The bucket size B is in days and is used whenever the analysis involves $\sigma(F)$ as this quantity is not defined for an individual query. The "origin dates" are used for the analysis in Section 5 to "align" queries.

epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[6] D. Gruhl, R. Guha, D. L. Nowell, and A. Tomkins. Information diffusion through blogspace. In *Conference on World Wide Web (WWW)*, pages 491–501, 2004.

[7] R. Jones, R. Kumar, B. Pang, and A. Tomkins. "I know what you did last summer": query logs and user privacy. In *Conference on information and knowledge management (CIKM)*, pages 909–914, 2007.

[8] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Conference on Knowledge discovery and data mining (KDD)*, pages 497–506, 2009.

[9] V. Mahajan, E. Muller, and F. M. Bass. New product diffusion models in marketing: A review and directions for research. *The Journal of Marketing*, 54(1):1–26, 1990.

[10] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Conference on Scalable information systems (INFOSCALE)*, page 1, 2006.

[11] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *Conference on Management of Data (SIGMOD)*, pages 131–142, 2004.

[12] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2003.

[13] I. Weber and C. Castillo. The demographics of web search. In *Conference on Research and development in information retrieval (SIGIR)*, pages 523–530, 2010.

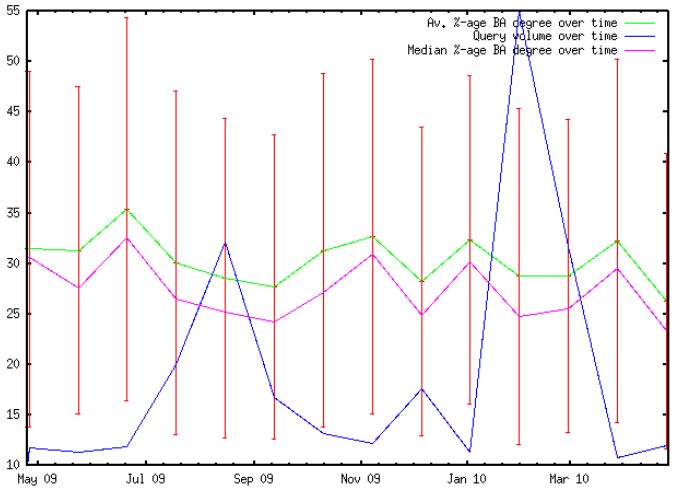

Figure 3: The query `ben bernanke` sees a drop in the feature "percentage holding a BA degree or higher" during times of increased search volume. This indicates that the users that normally issue this query have a higher education level than the typical users whose query appears to be driven by recent news events.