Characterizing Local Interests and Local Knowledge

Ryen W. White
Microsoft Research
Redmond, WA 98052
ryenw@microsoft.com

Georg Buscher
Microsoft Bing
Bellevue, WA 98004
georgbu@microsoft.com

ABSTRACT
When searching for destinations and activities, the interests and knowledge of locals and non-locals may vary. In this paper, we compare and contrast the search-related interests of these two groups, and when they share a common interest (in our case, for restaurants), we analyze the quality of the venues they intend to visit. We find differences in interests depending on local knowledge, and that locals generally select higher-quality venues than non-locals. These findings have implications for search and recommendation systems that can personalize results based on local knowledge and leverage that knowledge to benefit non-locals.

Author Keywords
Local interests; Local knowledge; Web search.

ACM Classification Keywords
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *search process*, *selection process*.

General Terms
Experimentation, Human Factors.

INTRODUCTION

A national US survey showed that people spend over two hours per day in everyday places in their vicinity, including restaurants, malls, and health clubs [5], affording ample opportunity to gain experiences that are useful in identifying candidate places, and distinguishing between places [6]. Since studies have shown that one quarter of Web search queries have local intent [3], people's local experiences may bring significant benefit to others searching for local information or afford search personalization based on search engine estimates of a user's local know-how.

We use the term *local knowledge* to describe an understanding of a particular location gained through experience with it. Despite the potential benefit of local knowledge to non-locals, it may be tacit and undocumented, and therefore challenging to derive practical value from. Large-scale Web logs gathered by search engine companies contain the search and browsing behavior of millions of users, including locals, and can implicitly reveal aspects of their local knowledge that are not documented online.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

Methods to support local search typically extract locations from queries and documents [1,4], but do not model local preferences based on a user's primary location. Other work has focused on location-wise personalization based on URL access patterns [2], but has not leveraged local knowledge. A better understanding of similarities and differences in local / non-local interests is lacking and necessary.

In this paper, we present a log study targeting two problems: (i) understand similarities and differences in local interests, and (ii) study differences in the real-world resources that locals select. We show that locals and non-locals have different interests about the same location, and that if we control for venue type (restaurants) locals make better decisions. These findings can inform the use of local knowledge for search support, including personalization.

RELATED WORK

There has been research on capturing and leveraging local knowledge. Wu et al. [10] mined Google *MyMaps* data and found that locals and non-locals referred to different landmarks in New York City when creating personalized city maps. Locals focused on daily life activities and newly-blooming neighborhoods whereas non-locals focused on tourist destinations and activities. The authors also present promising early findings from using collaborative filtering to generate recommendations via map co-occurrence. Ludford et al. [6] created a location-based reminder system, *PlaceMail*. The authors identified the heuristics people use when deciding which place information to share, how these findings relate to the design of local knowledge sharing systems, and to identify new uses of place information.

The information retrieval community has used using geographic criteria to retrieve documents [1]. Jones et al. [4] examined the effectiveness of geographic features of the document, the query, and the document-query combined, and trained a ranker to combine textual and geographic similarity (geo-spatial) features. Rather than mining locations and distances from Web page or query content, Bennett et al. [2] built models of the locations from which users view individual Web documents. They then personalized search result rankings based on both model properties and how typical the user's location is for each search result.

There has also been work on detecting and using locations. Mummidi and Krumm [8] leveraged users' map annotations to discover geographic points of interest. Mehler et al. [7] used locations mentioned in online news articles to detect regional biases toward entities such as players in local sports teams and local politicians.

We extend previous work in a number of ways: (i) we explore similarities and differences in *search-related* interests of locals and non-locals; (ii) we examine differences in the quality of the venues visited by locals and non-locals; and (iii) we present design implications for search and recommendation systems based on automatically inferring degrees of knowledge with respect to a location of interest.

UNDERSTANDING LOCAL INTERESTS

At the outset of our work, we wanted to understand similarities and differences between locals' and non-locals' search interests regarding the same location. To identify locals and non-locals automatically, we use log data containing the natural search behaviors of people in many locations.

Log Data

The primary source of data in this study is a proprietary data set comprising two months of anonymized logs (from February and March 2011) of URLs visited by users who consented to provide interaction data through a Web browser add-on widely distributed by the Microsoft Bing search engine. The logs comprised billions of queries and URL visits. We used the February data to identify locals (and non-locals), and the March data to study their search behavior. The data include a random unique user identifier, the date and time, and the URL of the Web page visited. Further, each user's IP address is resolved into geographic location information (i.e., city and state, plus latitude and longitude) and recorded. All log entries resolving to the same town or city were assigned the same latitude and longitude coordinates. To remove variability caused by cultural and linguistic variation in search behavior, we only include log entries from the English-speaking United States locale.

Identifying Locals

For each user in the February 2011 subset of our data, we computed a distribution of locations across all URL visits. To improve the reliability of the location distributions, we restricted our analysis to users with 100 or more URL visits over at least 14 days in the one-month period. Users for whom 100% of their page visits came from a single location were regarded as *locals*. Using this methodology, we extracted 4.5 million users who were local to a single location (88% of all users[1]). We identified locals for over 14,500 locations from all over the United States, comprising large cities (e.g., New York City (NYC), NY) and small cities (e.g., Woodinville, WA). Note that the *non-locals* for a given location were by definition locals in other locations.

Comparing Local and Non-Local Search Interests

From our March 2011 data we extracted search queries and URLs reached via search engine result page (SERP) clicks on the Google, Yahoo!, and Bing search engines. To help ensure we were dealing with local intent, we examined queries that contained a city, and the state name or state name abbreviation of the location of interest (e.g., *[hotels in abilene, texas]*). Local actions were identified as queries pertaining to the location of interest and their associated SERP clicks. A user's current location is less important in our analysis than their primary location. By using February data to identify locals, and the March data to compare local/non-local interests we allow for cases of travel in our analysis.

	URL domain or URL path	*Label*
Locals	**seattle.craigslist.org**	**Classifieds**
	metro.kingcounty.gov	Transit
	seattletimes.nwsource.com	**News**
	seattle.gov/light	Utilities
	komonews.com	News
	king5.com	**News**
	kingcounty.gov/healthservices	Health
	wsdot.wa.gov/traffic/seattle	Traffic
	zillow.com	Real estate
	costco.com	Commerce
Non-locals	**seattle.craigslist.org**	**Classifieds**
	en.wikipedia.org/wiki/seattle	Information
	king5.com	**News**
	seattle.gov	Government
	wunderground.com	Weather
	visitseattle.org	Tourism
	seattletimes.nwsource.com	**News**
	washington.edu	Education
	portseattle.org/seatac	Flights
	pikeplacemarket.org	Tourism

Table 1. Top-10 Seattle-oriented websites visited by locals and non-locals, plus topic labels. Bold = visited by both groups.

To measure the similarities and differences in interest between locals and non-locals, we computed ranked lists of URLs for each of the locations based on URL popularity with locals and non-locals. To understand the nature of the URLs selected, one of the authors also manually assigned topic labels (e.g., Classifieds, Tourism) to the URLs for a randomly selected set of 50 locations. The labeling scheme was iteratively refined as more URLs were encountered, and URLs re-labeled as necessary. Table 1 shows the top 10 URL domains (or URL paths, for clarity as needed) visited by locals and non-locals with interest in Seattle, WA.

Table 1 shows that Seattle locals were primarily interested in classifieds, news and traffic reports, transit, utilities, and (outside of the top-10 but also popular) hospitals and education (e.g., school districts). Non-locals were interested in classifieds, but also general information, news, travel, and tourism. This extends prior work [10], which showed similar differences in map labels assigned by NYC locals and non-locals, but did not study search interests.

Despite some differences, there were also common interests that were popular with both locals and non-locals, but since that interest did not center on a single URL, it is not represented in Table 1. One such case was restaurants; dining out is a popular interest in general, but that interest is spread

[1] To simplify the analysis we focused on the 88% of users in a single location during February 2011. The other 12% were mobile and although an analysis of the more mobile users is warranted, it is harder to identify them as locals.

	URL domain or URL path	Label
Non-locals: ≥ 1000km	**en.wikipedia.org/wiki/seattle**	**Information**
	seattle.gov/visiting	**Information**
	weather.com	Weather
	visitseattle.org	**Tourism**
	king5.com	News
	seattletimes.nwsource.com	Events
	graylineseattle.com	Buses
	seattlechildrens.org	Health
	swedish.org	Health
	portseattle.org/seaport	Cruises
Non-locals: ≥ 3000km	**seattle.gov/visiting**	**Information**
	en.wikipedia.org/wiki/seattle	**Information**
	visitseattle.org	**Tourism**
	farecompare.com	Flights
	seattle.apartments.com	Housing
	portseattle.org/seatac	Flights
	spaceneedle.com	Tourism
	pikeplacemarket.org	Tourism
	pioneersquare.com	Tourism
	seattlerecruiter.com	Jobs

Table 2. Variations in Seattle-oriented websites visited by non-locals whose primary locations were 1000km or 3000km from Seattle. Bold = visited by both groups.

over many different restaurant URLs. These commonalities prompted us to look more at how we identified non-locals.

Although Table 1 shows local / non-local differences, there are still popular non-local URLs that we would expect to only see locals visit (e.g., the local news site king5.com). One reason for this is our local / non-local labeling method: people only need to be in a different city to count as non-locals; we do not consider the *distance* from the location in our definition of non-local (people in a neighboring town or city are regarded as non-local irrespective of the distance between locations). To better account for distance from the originating city, we studied the interests of non-locals in two regions: (i) those at least 1000 kilometers (km) from Seattle (where it was not, reasonably, possible to drive to the location in a single day), and (ii) those at least 3000 km away (not, reasonably, drivable at all). Table 2 shows that the top URLs visited by non-locals varies with distance from location: 1000 km non-locals wanted activities, buses/cruises, and healthcare facilities; 3000 km non-locals wanted general information, flights, tourism, and information related to life changes (moving, jobs). From these findings it appears that when we hold location constant but vary the user group searching for that location, there are quite significant differences in their local interests, especially as the distance from the location increases. We applied the same 1000/3000 km thresholds to the other 49 US cities in our study (including a number of cities on the densely populated US East coast) and saw similar differences in local/non-local interests to those observed for Seattle.

Although locals and non-locals may have different interests when searching for a location, as noted earlier there are common interests in topics such as dining out. Studying shared interests affords a direct comparison of the decisions made by local and non-locals while controlling for the domain of interest. Since dining out is such a popular activity we explore differences in the quality (and nature) of the restaurants that locals and non-locals intend to visit.

LOCAL KNOWLEDGE IN PRACTICE

We first describe our procedure to estimate *restaurant visitation intent* (RVI) and obtain restaurant quality ratings.

Estimating Restaurant Visitation Intent

1. Extract sessions from the March 2011 log data described earlier using an approach similar to [9]. Browse sessions begin with a user opening the Web browser window, and end with an inactivity timeout of 30 or more minutes.
2. Find instances of users making restaurant reservations. To do this we automatically search sessions for evidence of the OpenTable (opentable.com) reservation site URL, which contain a distinct URL pattern for reservations.
3. Automatically classify URLs in a session as belonging to the user's primary location (hereafter referred to as "local URLs") using a proprietary classifier with features such as addresses in page content, addresses in queries leading to SERP clicks on those pages, etc.
4. Given the occurrence of a reservation in a session and at least one local URL preceding the reservation, we assume that the most proximal prior local URL to the reservation URL was a restaurant of interest, and that the (user, location, restaurant)-tuple represents an RVI[2]. Visual inspection of the sessions showed that this method correctly identified restaurants in most cases. URLs incorrectly labeled with this approach included were for review sites, theatres, hotels, resorts/spas, wineries, and chocolatiers. Those erroneous RVIs were excluded.

Using the expert identification method described in the previous subsection we identified instances of locals and non-locals with RVIs for the same location. To compare the quality of restaurants that locals and non-locals reserved, we searched the following five popular restaurant review sites for the restaurant and the location of interest: OpenTable, Urbanspoon (urbanspoon.com), Yahoo! Local (local.yahoo.com), TripAdvisor (tripadvisor.com), and Yelp (yelp.com). For each restaurant, we averaged ratings across all sites with reviews of it and obtained a final rating from 1 to 5. In total, there were 1,267 RVI instances at 984 distinct restaurants, each with an aggregate quality rating.

Comparing Local and Non-Local Restaurant Selections

Table 3 shows the mean (M) and standard deviation (SD) rating. N is the total number of RVIs in each group. As described earlier, we also examined the distance of non-locals from the restaurant and filtered to non-locals at least 1000 km or 3000 km away from the restaurant city.

[2] We assume that reserving a table at a restaurant is a reasonable surrogate for an RVI. However, given the nature of our data we could not confirm this. Also note that we could not determine the date or time of the intended visit, or the restaurant itself, directly from the reservation URL.

	Local RVIs	*Non-local RVIs*		
		All	Distance from restaurant (city)	
			≥ 1000 km	≥ 3000 km
M	4.00	3.89	3.87	3.83
SD	0.55	0.59	0.54	0.56
N	233	661	141	43

Table 3. Average, standard deviation, and number of ratings assigned across all restaurants visited by locals and non-locals. Rating scale is 1 to 5, higher is better.

Table 3 shows that there are differences in the ratings assigned to restaurants visited by locals and non-locals. Unpaired t-tests (with $\alpha = 0.05$) between the average ratings assigned to the places that locals and non-locals visit revealed that the differences were significant ($t(982)=2.33$, $p=0.02$). The findings also show that non-locals reserved tables at lower quality restaurants (1000 km: $t(372)=2.58$, $p=0.01$; 3000 km: $t(274)=2.59$, $p=0.01$). Although the rating differences may appear small, we average over reviews from different sites to obtain a wide range of opinions. We believe that the trend in the findings (i.e., restaurant ratings drop with non-locals and increased distance) is noteworthy.

To better understand the nature of differences in the restaurants locals and non-locals visited, we studied additional features. We randomly selected a set of 125 restaurants from Seattle and ten other US cities, all of which had local and non-local restaurant visits, and for each restaurant we visited OpenTable to obtain price level (on a scale from 1 to 4), cuisine, and whether the restaurant was in a hotel (and hence potentially convenient for non-locals). Our findings show that non-locals selected slightly cheaper restaurants (local=2.64, non-local=2.52)—although not significantly so (Mann-Whitney test, $U(125)=1783$, $p=0.32$)—and exhibited different cuisine preferences (e.g., locals in Houston, TX seemed to prefer steakhouses).

One explanation for locals intending to visit higher-rated restaurants might be that it is mainly locals who are providing the ratings, and reflecting their own experiences with the restaurants. However, from examining the review sites used in our analysis, we see that ratings are provided by a broad mixture of locals and non-locals. The differences may in part relate to traveling users (e.g., a NYC local dining out in Seattle) being more constrained in their activities. However, we did not observe differences between our local and non-local groups in the frequency with which group members visited restaurants in hotels.

DISCUSSION AND IMPLICATIONS

We showed that there were commonalities and differences in interests between locals and non-locals, and that these differences were more pronounced when we included distance. Locals selected better quality restaurants, and there were indications of differences in price and cuisine preferences between locals and non-locals. The implications of these findings fall into two main areas:

Personalizing to local interests: Our analysis showed that locals and non-locals seek different information about the same location. Search and recommendation systems could personalize based on whether a user is a local, perhaps by applying a ranking algorithm giving differential weight to tourist sites. Since we also found differences in users' interests per their distance from the target location, search systems could also leverage distance between a user's primary (not necessarily current) location and the target as a ranking feature or as a trigger for showing local event information or social recommendations (e.g., local friends' suggestions).

Leveraging local knowledge: The lower quality ratings for restaurants non-locals intend to visit underscore the need for better support for non-locals' selection of local venues and activities. To help, we could highlight local favorites to non-locals directly on SERPs for local queries or leverage the search behavior of locals mined from log data to improve the quality of the results returned for local queries.

Beyond technological augmentations, there are also important social implications from leveraging local knowledge that must be considered. For example, directing non-locals to popular local attractions may turn local gems into tourist hotspots, detracting from their quality. Qualitative data on local interests and knowledge are also needed to complement the quantitative analysis described in this paper. In future work we will explore local knowledge in more detail, implement local-knowledge-based search personalization and develop search support that uses locals' search behavior to benefit non-locals, and evaluate our system enhancements via user studies and Web-scale deployments.

REFERENCES

1. Amitay, E. et al. Web-a-where: Geo-tagging web content. *SIGIR*, (2004), 273–280.
2. Bennett, P.N. et al. Inferring and using location metadata to personalize Web search. *SIGIR*, (2011), 135–144.
3. Himmelstein, M. Local search: The internet is the Yellow Pages. *IEEE Computing 38*, 2, (2005), 26–34.
4. Jones, R., Hassan, A. and Diaz, F. Geographic features in Web search retrieval. *ACM GIR Workshop on Geographic Information Retrieval*, (2008), 57–58.
5. Klepis, N. et al. The national human activity pattern survey. *J. Exposure Analysis and Environmental Epidemiology 11*, 3, (2001), 231–252.
6. Ludford, P.J. et al. Capturing, sharing, and using local place information. *SIGCHI*, (2007), 1235–1244.
7. Mehler, A. et al. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics 12*, 5, (2006), 765–772.
8. Mummidi, L. and Krumm, J. Discovering points of interest from users' map annotations. *GeoJournal 72*, (2008), 215–227.
9. White, R.W. and Drucker, S.M. Investigating behavioral variability in Web search. *WWW*, (2007), 21–30.
10. Wu, S. et al. Mining collective knowledge from Google MyMaps. *WWW*, (2011), 151–152.