Augmenting Business Entities with Salient Terms from Twitter

Riham Mansour
Microsoft Research ATL
Cairo, Egypt
rihamma@microsoft.com

Nesma Refaei
Cairo University
Cairo, Egypt
nesma.a.refaei@eng.cu.edu.eg

Vanessa Murdock
Microsoft
Seattle, WA
vanessa.murdock@yahoo.com

Abstract

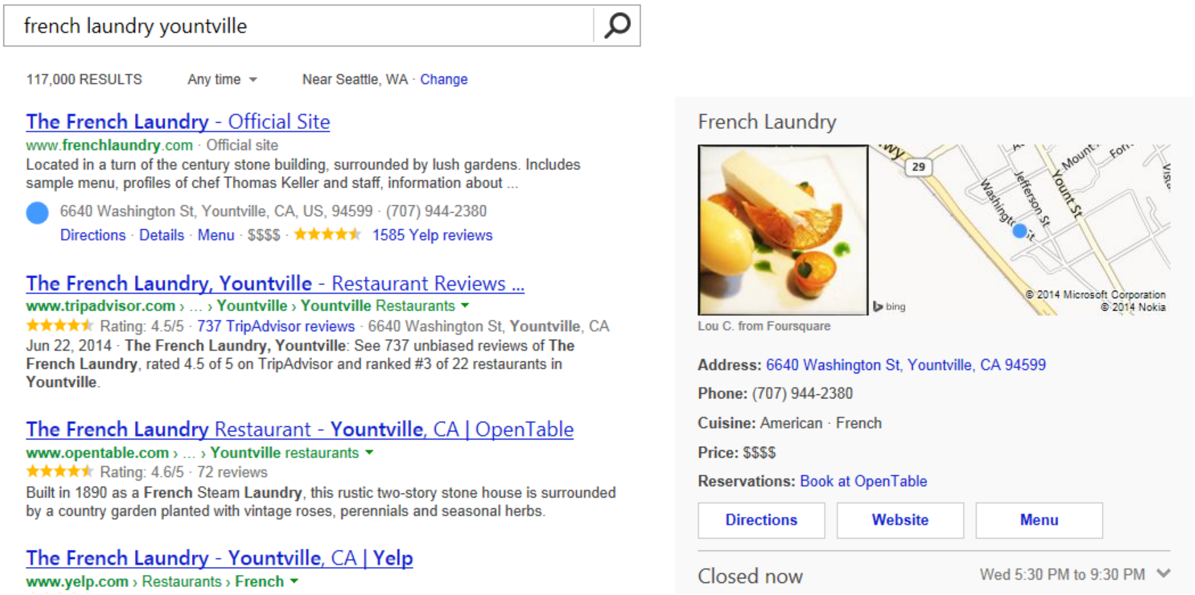
A significant portion of search engine queries mention business entities such as restaurants, cinemas, banks, and other places of interest. These queries are commonly known as "local search" queries, because they represent an information need about a place, often a place local to the user. A portion of these queries is not well served by the search engine because there is a mismatch between the query terms, and the terms representing the local business entity in the index. Business entities are frequently represented by their name, the category of entity (whether it is a restaurant, an airport, a grocery store, etc.) and other meta-data such as opening hours and price ranges. In this paper, we propose a method for representing business entities with a term distribution generated from web data and from social media that more closely aligns with user search query terms. We evaluate our system with the local search task of ranking businesses given a query, in both the U.S. and in Brazil. We show that augmenting entities with salient terms from social media and the Web improves precision at rank one for the U.S. by 18%, and for Brazil by 9% over a competitive baseline. For precision at rank three, the improvement for the U.S. is 19%, and for Brazil 15%.

1 Introduction

Search engine queries, particularly queries issued from mobile devices, often mention business entities such as restaurants, cinemas, banks, and other places of interest. These "local search" queries represent an information need about a place. Often there is a mismatch between the query terms, and the terms representing the local business entity in the index, making it difficult for the search engine to find results that satisfy the user. Local data consists largely of listings of businesses, annotated with metadata. This metadata includes the name of the location, category information (is the business a clothing retailer, or a Thai restaurant, for example), address and phone number, opening hours, and indicators such as price range, popularity, star ratings, etc. Figure 1 shows an example of the type of information available to local search systems.

Some local search queries are known item searches, where the user knows the name of a business and they seek other information about the place, such as the opening hours. Other local search queries are category searches where the user does not know the name of a specific business but is using the Internet in much the same way they might have used the Yellow Pages in pre-Internet days. An example of a category search is "Thai restaurants in Denver". There are also descriptive local queries such as "pizza delivery" or "romantic brunch in Seattle" where the user does not mention a category or a business name directly, but for which there is a closed class of businesses that will satisfy the user's need.

Descriptive queries such as "roasted chiles in Santa Fe" or "kid-friendly Caribbean resorts" pose a significant challenge to local search systems, as the information in the local index does not typically include terms that match the user's query. That is, the system may know businesses in Santa Fe, but not whether they sell roasted chiles.



french laundry yountville

117,000 RESULTS Any time Near Seattle, WA · Change

The French Laundry - Official Site
www.frenchlaundry.com · Official site
Located in a turn of the century stone building, surrounded by lush gardens. Includes sample menu, profiles of chef Thomas Keller and staff, information about ...
6640 Washington St, Yountville, CA, US, 94599 · (707) 944-2380
Directions · Details · Menu · $$$$ · ★★★★★ 1585 Yelp reviews

The French Laundry, Yountville - Restaurant Reviews ...
www.tripadvisor.com › ... › Yountville › Yountville Restaurants
★★★★★ Rating: 4.5/5 · 737 TripAdvisor reviews · 6640 Washington St, Yountville, CA
Jun 22, 2014 · The French Laundry, Yountville: See 737 unbiased reviews of The French Laundry, rated 4.5 of 5 on TripAdvisor and ranked #3 of 22 restaurants in Yountville.

The French Laundry Restaurant - Yountville, CA | OpenTable
www.opentable.com › ... › Yountville restaurants
★★★★★ Rating: 4.6/5 · 72 reviews
Built in 1890 as a French Steam Laundry, this rustic two-story stone house is surrounded by a country garden planted with vintage roses, perennials and seasonal herbs.

The French Laundry - Yountville, CA | Yelp
www.yelp.com › Restaurants › French

French Laundry

Lou C. from Foursquare

Address: 6640 Washington St, Yountville, CA 94599
Phone: (707) 944-2380
Cuisine: American · French
Price: $$$$
Reservations: Book at OpenTable

Directions Website Menu

Closed now Wed 5:30 PM to 9:30 PM

Figure 1: Example of the type of meta-data associated with a business entity, in this case a restaurant.

However, collectively people themselves know this type of information, and they frequently mention it in social media. The discussion of a local business in social media, such as Twitter[1], Flickr[2], Facebook[3] and Foursquare[4] may take the form of a simple check-in ("Drinking a Smog Rocket at @byronhamburgers") or a Facebook status caption to a photo ("Sea stars at the Seattle Aquarium"), or a Tweet ("the quad & the blonde both were good! The choc flavored one wasn't so much to my tastes..."), among others.

A growing number of users of social media attach geographic coordinates to their status updates, allowing the text of the updates to be associated to a location. Further, businesses use social networks as a publicity platform to widen their customer base. Today, Twitter has more than 500 million users.[5]

In this paper we augment business entities with salient terms describing the business. We extract the terms from Twitter, and from the Web. To determine which terms are salient, we compute the co-occurrence of terms with mentions of the business name (and name variants), for tweets issued within one kilometer of the business. Because some users are especially prolific on social media, and may dominate the tweets issued in that location, we estimate the term co-occurrence statistics with the user frequency of a term: the number of people using that term in a given location. We also extract salient terms from the Web pages of the business entity, but in this case the user frequency is not meaningful, so term co-occurrence is calculated with the term frequency.

We evaluate the term distributions describing a place in the context of local search for the U.S. and Brazil. We construct a corpus of search engine queries with local intent, and evaluate the retrieval of businesses in response to the queries. We compare several different strategies for augmenting the representation of the business to a baseline system described in Colombo et al. (2013). Augmenting with tweets improves precision at rank one for U.S. local search by 18%, and for Brazil by 9%. For precision at rank three, the improvement for the U.S. is 19%, and for Brazil 15%.

The rest of the paper is organized as follows: Section 2 surveys the related work. Section 3 details how salient terms are extracted from tweets and the Web. Section 4 illustrates the experimental setting and the evaluation of the impact of salient terms on retrieval. Section 5 presents a discussion of the results and Section 6 concludes the paper with remarks for future work.

[1] www.twitter.com visited March 2014
[2] www.flickr.com visited March 2014
[3] www.facebook.com visited March 2014
[4] www.foursquare.com visited March 2014
[5] http://www.statisticbrain.com/twitter-statistics/ visited March 2014

2 Related Work

Modeling business entities from multiple sources like the Web and social media remains an open problem. Most of the work in this domain focuses on modeling locations and regions more generally (O'Hare and Murdock, 2013; Laere et al., 2012; Bennett et al., 2011), or on extracting mentions of business entities from text using NLP techniques (Rae et al., 2012). O'Hare and Murdock (2013) propose a statistical language modeling approach to characterize locations in text, based on user frequency. They utilize the geo-tagged public photos in Flickr. The primary difference between their work and ours is that they estimate the user frequency distribution, whereas we employ the user frequency in calculating the term co-occurrence. Also, the locations described in O'Hare and Murdock represent locations of one kilometer distance. They do not attempt to characterize specific points of interest or businesses.

There has been significant effort to leverage image content to characterize locations, due to the availability of geotagged Flickr photos. Much of the work uses Flickr photos and tag sets, and focuses on identifying the locations in photos. This is related, although not directly applicable, to work with Twitter. Ahern et al. (2007) identify geographically related tags by finding dense areas using geodesic distances between images. They rank the tags in these areas with $tf.idf$. In their subsequent work Kennedy et al. (2007) leverage tags that represent local events. Naaman et al. (2003) and Moxley et al. (2008) propose approaches for recommending tags to the user given a known location for an image. Some research efforts leverage image content to characterize locations. Crandall et al. (2009) employs image content and textual metadata to predict the location of a photograph at the city level and at the individual landmark level. Hays and Efros (2008) use visual features to predict geographic locations by nearest-neighbor classification.

Colombo et al. (2013) provide the baseline system for this paper, and it is described in more detail in Section 4.1. They use online reviews, comments and user tips about points of interest in location-based services like Yelp, Google+, and Qype to build a tag-based representation of a point of interest. They rank the tags by their $tf.idf$ score from a collection of location-based service related documents.

In terms of using geo-referenced information to represent locations, Rodrigues (2010) proposes to extract points of interest automatically from the Web, for example from Yahoo, Manta and Yellow Pages. He also infers points of interest based on geo-referenced content such as geo-tagged photos, blog posts and news feeds. They cluster content from multiple sources while building a language model for each cluster. Tags in each cluster are scored by $tf.idf$. This work is similar in spirit to the work proposed in this paper, although our work focuses more on obtaining the most unique and frequent tags associated with points of interest in tweets.

Hegde et al. (2013) assign tags to points of interest based on user interest profiles in online social networks and check-in logs of users at these places. They use probabilistic modeling to derive the point of interest tags followed by hierarchical clustering of most probable tags to filter out semantically irrelevant tags. Biancalana et al. (2013) use point of interest-related location-based service content to extract key phrases that could serve as tags characterizing each point of interest. The extracted phrases are weighted by user authority.

In terms of modeling locations from short microblog messages like tweets. Paradesi (2011) proposes TwitterTagger, a system that geo-tags tweets and shows them to users based on their current physical location. The tweets are geo-tagged by identifying the locations referenced in a tweet by part of speech tagging and a database of locations. Eisenstein et al. (2010) and Kinsella et al. (2011) present methods to identify the location of a user based on his or her tweets. Li et al. (2011) rank a set of candidate points of interest using language and temporal models. Given a query tweet, they build a unigram language model for each candidate point of interest and for the query tweet. Points of interest are then ranked by their KL-divergences with the tweet language model. Unlike our work, both approaches identify a location in tweets rather than modeling a certain location by the way it is mentioned in tweets.

3 Describing Businesses with Twitter and the Web

Salient terms are terms that uniquely characterize a place. As an overview, we extract terms from two sources namely geo-tagged tweets and business-related webpages. We extract terms from geo-tagged tweets posted from locations within one kilometer of the business. We then identify the tweets about a given business from among the nearby tweets, by looking for mentions of the business name (along with naming variants). We compute the term co-occurrence between the business name, and the terms that occur in tweets mentioning the business.

We also extract terms from webpages related to the business entity. We issue a query with the business name to the Bing Search API.[6] We compute the term co-occurrence between the business name, and the terms that occur in these top three web pages resulting from Bing search.

There is no universal standard for representing locations. Some gazetteers are available for developers that represent places according to a hierarchy (such as Geonames[7] and Placemaker[8]). There is also proprietary data gathered by companies such as Nokia, YellowPages and Yelp, which provide some information about places like geo-location, address, and phone number. There are also open source data like Freebase and DBpedia. Both proprietary and open source data use structured representations for places.

There are three challenges with these representations. First, they do not provide a rich description of the place, as they are primarily designed to help users locate the place, via the name, address and phone number, or category ("restaurant" or "cinema," for example). However, the categories may be broad and in a language different from the language spoken by the user. Second, the coverage of points of interest and businesses focuses mostly on well-known places. Businesses are not usually well-represented because they are often relatively ephemeral. Finally, the data may be stale. For example, a restaurant that has closed, or moves location, should be flagged, and it may take time for the gazetteer to be updated. Social media provides fresh information about businesses, especially as more businesses promote themselves via these channels. Modeling businesses with tweets could complement the available data with fresh descriptions.

3.1 Text Pre-processing

We acquire geo-tagged tweets related to business entities in the United States and Brazil from the Twitter firehose, from January 1, 2013 to May 31, 2013. We chose these countries because of their high usage of Twitter, and to show that the approach is language agnostic. The tweets are primarily in English (in the U.S.) and in Portuguese (in Brazil).

We pre-process the tweets by removing stop words, using the Natural Language Toolkit (NLTK) library[9] and non-alphabetic characters. For our baseline implementation following Colombo (2013), we remove the non-English words using the English NLTK wordnet corpus. For removing non-Portuguese words in the baseline, we use the Enchant specll checking library.[10] In our proposed approach, we don't remove non-English and non-Portuguese words, but we rather remove twitter terms that did not appear in the Bing query logs in December 2013. Further, we remove tweets automatically generated by check-in services such as Foursquare by detecting the patterns "I'm at" and "mayor". We remove shortened URLs in the tweet text by detecting the pattern "http://t.co." URLs are removed as they do not carry salient terms. All text was lower-cased. All tweets are indexed in Solr,[11] an open-source search engine which allows for field search. The index carries the tweet text, geographic coordinates, time stamp, language, country, retweet count, source, URL and user information.

3.2 Computing Salient Terms

The business entities were submitted to the Solr index as queries, to retrieve the tweets related to the entity itself. We apply two sequential filters on the indexed tweets to obtain the relevant tweets. The first filter limits the search to those tweets whose geographic coordinates are within one kilometer of the business entity. This covers a wide range around the POI due to the small volume of geo-tagged tweets

[6] http://datamarket.azure.com/dataset/bing/search visited March 2014

[7] http://www.geonames.org visited March 2014

[8] http://developer.yahoo.com/boss/geo/ visited March 2014

[9] http://nltk.org/ visited March 2014

[10] http://pythonhosted.org/pyenchant/ visited March 2014

[11] http://lucene.apache.org/solr/ visited March 2014

in general. Enlarging the range to one kilometer retains a reasonable volume although it does introduce more irrelevant tweets. The second filter eliminates irrelevant tweets by searching with the canonical name of the business along with naming variants. The indexed tweets are searched by name, 70% of the name, and the name fully concatenated with no spaces separating the multiple words, and with spaces replaced with an underscore. The resulting set of tweets are those that are relevant to the business entity since they have been posted within its vicinity and they mention the entity directly.

To extract the salient terms from Twitter, we compute the term co-occurrence of the entity name with the set of terms co-occurring in the associated tweets. Term co-occurrence is traditionally computed as the number of times term t and term w appear in the same tweet C, divided by the number of times term t appears in any tweet in the same one-kilometer vicinity, plus the number of times term w appears in any tweet in the same one-kilometer vicinity:

$$score(t,w) = \frac{count_C(t,w)}{count_C(t) + count_C(w)}. \tag{1}$$

Some users of twitter are extremely prolific, and may generate a lot of data in a small set of places. Term frequency may produce an estimate of the term distribution biased toward a particular user or set of users. To prevent a single prolific user from dominating the representation of a place, we estimate the term co-occurrence with the user frequency. That is, the term counts are the number of people who used a term in a place, rather than the number of times a term was applied. This has been shown to be a more reliable estimate of term distributions in other work using social media to model places (O'Hare and Murdock, 2013). Note that the baseline implementation is based on the term frequency, and uses tf.idf rather than term co-occurrence.

We also enrich the business entities with terms from the web pages. We issue a query to Bing Search API with the business name. We then extract salient terms from the content of the top three results. We pre-process the text according to Section 3.1 to get the unigram terms. We filter out the terms that are substrings of the business name, and single character terms. The terms are weighted according to the term frequency (tf) and the terms with $tf > 0.001$ are considered salient to the business entity. This threshold has been selected empirically.

4 Experimental Setting

In our experiments we evaluated the effect of expanding the business entities with salient terms within the context of local search. We examined whether adding tags such as "conchiglie" to the entity "French Laundry" will improve the retrieval results for a query with local intent like "conchiglie Napa Valley". For this purpose, we sampled a set of 30,000 businesses from a proprietary database of business listings in the United States and Brazil. We then chose 80 entities from the two countries to formulate the test set of search queries as illustrated below.

4.1 Baseline Approach

Colombo et al. (2013) suggested a method for filtering the salient terms extracted from a set of documents relevant to a place of interest. We used their method to filter the salient terms extracted from the geo-tagged tweets selected and pre-processed as described above in Section 3.1. The terms remaining after these filtration steps are weighted using $tf.idf$, where a background corpus of all tweets relating to any business within one kilometer of the entity in question is used to calculate the idf of each term. Finally, we kept only the terms with a $tf.idf$ greater than a threshold of 0.04 as the baseline salient terms for the business.

4.2 Building the Search Corpus

Our database of businesses contains metadata about each business including the name, phone number, website, street address, city, country, geographic coordinates, and category information that are a subset of a taxonomy of categories both in English and in the language of the country of the business. We appended the extracted salient terms for each business as a field in our database. We removed twitter terms that escaped initial filtering by removing any terms that did not appear in the Bing query logs in December 2013. We also filtered out twitter terms that are included in the category taxonomy, as these tags will not add value to the existing data, and are unlikely to improve retrieval over the naive baseline.

Some businesses are very popular, and are likely to generate more social media traffic. To make sure that the system is as general as possible, and that we don't build in an inherent bias toward popular businesses (or national chains) we construct the search corpus to represent varying popularity levels. The popularity of a business is quantified by the number of unique users tweeting about it. We stratify the selection of the businesses from our database of 30000 businesses such that the search corpus contains 15,000 businesses from the U.S. and 15,000 businesses in Brazil, which are distributed across a range of popularity scores. Finally we indexed the search corpus using Solr.

4.3 Generating Search Queries

We formulate search queries by selecting 40 businesses in each market with their attributes and salient terms. We formulated query templates from the business name, location, category and terms selected by three judges from associated tweets and Web pages. The information is detailed in Table 1. The query templates are shown in Table 2, along with an illustrative example of each one.

Attribute	**Description**
Name	business name and variants
Location	city and country
Categories	categories provided by the database
Terms	term selected by judges from Twitter and Web pages

Table 1: Information included in the baseline queries

Query Template	**Example**
Name	"French Laundry"
Name + Location	"French Laundry in Yountville" or "French Laundry in California"
Name + Category	"French Laundry Restaurant"
Name + Term	"conchiglie French Laundry"
Term + location	"conchiglie Yountville" or "conchiglie California"
Category + location	"Restaurants in Yountville" or "Restaurants in California"

Table 2: Query templates with examples

Some of the automatically generated queries (such as "happy in california" and "week in Houston") don't have a local intent because of uninformative terms (such as "good", "happy", or "week") or because of malformed substrings of names and categories. To filter out these uninformative queries we issued the query to Bing Search API and kept only the queries that generated a direct answer. An example of a direct answer is shown in Figure 1. The Bing Search API returns a direct answer when the query has been classified as having local intent. We use the Bing API in this way as a black box, because building a local intent classifier is a significant undertaking, and is beyond the scope of this paper. The resulting test set consists of 1000 local queries representing 80 business entities in Brazil and the U.S., with an equal distribution of each of the query templates in Table 2.

4.4 Evaluation

Our primary evaluation is of query expansion for the class of queries for which a business listing is a relevant result. However, representing a business entity with a term distribution estimated from social media has other applications as well. For this reason, we would like to know the quality of the expansion terms, independent of any task. To this end, we asked three judges to pick all the relevant terms from among an unordered set of extracted terms salient to a business, for 100 businesses in each country. We divided the terms among the three judges equally and each term has been judged by only one judge. The number of tags extracted from the web pages is an order of magnitude larger than the number of tags extracted from Twitter for a given business. We consider the tag accuracy to be proportion of "good" tags accounted for by a single data source. That is, for Twitter, it is the number of "good" Twitter tags, divided by the total number of "good" tags, whereas the accuracy of the Web tags is the number of "good" tags derived from the web, divided by the total number of "good" tags. Based on this assessment, the accuracy of the Twitter tags for the U.S. data was 0.22, and the accuracy of the Web tags was 0.78. For the data from Brazil, the accuracy of the Twitter terms was 0.15, and the accuracy of terms derived from the Web was 0.85.

The effect of the expansion strategies on the retrieval of business entities. As Solr allows for field search, we can limit the fields to the entity and its metadata, or the entity metadata and the twitter tags, etc. Tables 3 and 4 show the results for various retrieval from fields representing document expansion strategies on data from the U.S. and Brazil, respectively. The results are averaged over 500 queries (from the query formulations described above) for each country. In Tables 3 and 4 we see that nearly 60% of queries return the correct result at rank one, when the entity is represented only by its metadata. The results reported in the other rows also include the entity metadata. (The baseline in Tables 3 and 4 is described in Section 4.1.) Expanding the represnetation of the point of interest with terms from the Web and from social media shows a clear benefit.

Mobile devices are becoming ubiquitous, and local search represents an important class of search on mobile devices. Because the devices are small, real estate to show results is extremely limited. For this reason, we choose to evaluate precision @ k, for $k <= 3$ for this task. To create a truth set, the top three results were evaluated by judges to determine their relevance to the query. Each result is judged by one assessor. Because precision at one is binary, we do not apply a statistical significance test. Percent change is reported for precision at rank one, with respect to the baseline (row two). The fact that the precision at rank three is lower than precision at rank one is an artifact of their being a single relevant result in most cases.

	P@1	**P@3**	**% Change in P@1**
Entity metadata	0.595	0.353	(oracle)
Baseline	0.627	0.358	NA
Entity metadata + twitter tags	0.667	0.389	+6.4%
Entity metadata + web terms	0.686	0.396	+9.4%
Entity metadata + web terms + twitter tags	0.738	0.425	+18%

Table 3: Precision @ k for local search in the U.S.

	P@1	**P@3**	**% Change in P@1**
Entity metadata	0.618	0.436	(oracle)
Baseline	0.643	0.460	NA
Entity metadata + twitter tags	0.650	0.474	+1%
Entity metadata + web terms	0.700	0.517	+8.9%
Entity metadata + web terms + twitter tags	0.708	0.533	+10%

Table 4: Precision @ k for local search in the Brazil.

5 Discussion

Since the set of queries consists of the entity name plus attributes from the index such as the location and the category information, the resulting precision from search just on the entity metadata itself shows the degree to which the bias in the data accounts for the results. That is, if you have the correct entity name, location and category, just searching for a business with matching metadata gives a precision at rank one of 0.595 (0.618 for Brazil). This is a naive baseline. The baseline results show that it is a competitive baseline because it demonstrates that there is a benefit to expand the representation of a business entity with text, beyond the naive baseline above it in the table.

The gains in precision suggest that the extracted salient terms with co-occurrence statistics and user frequency from twitter and the web pages are of better quality than the terms extracted by the baseline in Colombo et al. (2013) with term frequency only. This is attributed to the fact that co-occurrence statistics and user frequency capture the terms that people frequently use when describing a place. Further, the quality of the salient terms extracted from the web pages exceeds the quality of the twitter terms. This is to be expected if the main search results for a business entity are reasonable, and the top three results are relevant to the query. Social media is notoriously noisy, so it is not surprising that the web pages produce more reliable expansion terms. Furthermore, comparing the terms expanded from the web, to the terms expanded from Twitter, we see the relative improvement with respect to the baseline of the Web expansion terms is greater than the Twitter expansion terms. The fact that both expanding from twitter and the Web produces results better than either individually shows that the two term distributions cover different slices of the vocabulary.

We experimented with the number of tweets required to improve the representation of the point of interest. We focused on the portion of the test set with queries of the form *term + location* like "conchiglie Yountville," as those are the queries that are not answered with relevant results in the absence of the proper salient terms. We found that 10 to 30 tweets mentioning the business were sufficient to improve the retrieval results for these queries, and there was no benefit to increasing the number of tweets to 50 or 100. In the Brazil data, the results for four of the queries of the form *term + location* were degraded when sampling terms from 10 tweets compared to more. However, the results were the same for 30, 50, 100 or more tweets, suggesting that there is no benefit to increasing the number of tweets beyond 30. This suggests that a smaller number of tweets is better, in terms of extracting salient terms. One possible reason for this is that adding more tweets increases the number of noise terms, relative to the number of salient terms.

6 Conclusion and Future Work

In this paper, we present an effective representation of business entities with a term distribution generated from web data and from social media that more closely aligns with user search query terms. We evaluate our system with the local search task of ranking businesses given a query, in both the U.S. and in Brazil. Our method uses co-occurrence statistics and user frequency to extract relevant salient terms. The results demonstrate the effectiveness of this approach when compared with a competitive baseline that uses term frequency to extract salient terms. Furthermore, we show that query expansion with salient terms improves retrieval in the common task of retrieving a business listing in response to a user query.

We leave to future work applying query expansion from social media to larger collections of local search queries, and other methods for formulating query templates based on the metadata available with business listings.

References

Shane Ahern, Mor Naaman, Rahul Nair, and Jeannie Yang. 2007. World Explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07*.

Paul N. Bennett, Filip Radlinski, Ryen W. White, and Emine Yilmaz. 2011. Inferring and using location metadata to personalize web search. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*.

C. Biancalana, F. Gasparetti, A. Micarelli, and G. Sansonetti. 2013. An approach to social recommendation for context-aware mobile services. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1).

G. B. Colombo, M. J. Chorley, V. Tanasescu, S. M. Allen, C. B. Jones, and R. M. Whitaker. 2013. Will you like this place? a tag-based place representation approach. In *International Workshop on the Impact of Human Mobility in Pervasive Systems and Applications*.

D.J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*, pages 761–770. ACM.

Jacob Eisenstein, Brendan O'Connor, Noah Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP*.

James Hays and Alexei A. Efros. 2008. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Vinod Hegde, Josiane Xavier Parreira, and Manfred Hauswirth. 2013. Semantic tagging of places based on user interest profiles from online social networks. In *Advances in Information Retrieval: Lecture Notes in Computer Science*, volume 7814, pages 218–229. Springer.

Lyndon Kennedy, Mor Naaman, Share Ahern, Rahul Nair, and Tye Rattenbury. 2007. How flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of the 15th International Conference on Multimedia*, pages 631–640.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'm Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Content*, pages 61–68.

Olivier Van Laere, Steven Schockaert, and Barth Dhoedt. 2012. Georeferencing flickr photos using language models at different levels of granularity: An evidence based approach. *Journal of Web Semantics*, 16.

Wen Li, Pavel Serdyukov, Arjen de Vries, Carsten Eickhoff, and Martha Larson. 2011. The where in the tweet. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM)*.

Emily Moxley, Jim Kleban, and B.S. Manjunath. 2008. Sprititagger: A geo-aware tag suggestion tool minded from flickr. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR'08)*, pages 24–30.

Mor Naaman, Andreas Paepcke, and Hector Garcia-Molina. 2003. From where to what: metadata sharing for digital photographs with geographic coordinates. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 196–217.

Neil O'Hare and Vanessa Murdock. 2013. Modeling locations with social media. *Journal of Information Retrieval*, 16(1).

Sharon Paradesi. 2011. Geotagging tweets using their content. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*.

Adam Rae, Vanessa Murdock, Adrian Popescu, and Hugues Bouchard. 2012. Mining the web for points of interest. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Filipe Rodrigues. 2010. *POI Mining and Generation*. Ph.D. thesis, University of Coimbra.