Understanding Travel from Web Queries using Domain Knowledge from Wikipedia

Chander J Iyer
chander.iyer@verizonmedia.com
Yahoo! Research
Sunnyvale, California

Srinath Ravindran
rsrinath@verizonmedia.com
Yahoo! Research
Sunnyvale, California

ABSTRACT

Developing a deeper understanding of the travel domain is helpful for presenting users with consistent and reliable information, and few sources of data are able to achieve that. Further, such information can serve as background knowledge for evaluating machine learning algorithms. In this paper, we present part of our work towards developing such an understanding. We demonstrate a simple extraction technique and how the extracted data can be used to evaluate an unsupervised embedding model built on search queries with travel intent.

CCS CONCEPTS

• **Computing methodologies → Cluster analysis**; • **Information systems →** *Language models*; Wikis.

KEYWORDS

Wikipedia; label generation; model evaluation; query embedding

ACM Reference Format:
Chander J Iyer and Srinath Ravindran. 2019. Understanding Travel from Web Queries using Domain Knowledge from Wikipedia. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3316759

1 INTRODUCTION

Planning a trip is one of the many stressful things we do everyday, and lack of a good source of information is one of the major reasons. Travel typically constitutes a trip to a destination including the mode of transport to the destination, the means of stay and commute within the destination. Various websites such as Tripadvisor, Expedia, and other travel blogs provide the reviews and details about hotels, flights, and car reservations. However, they either (a) lack consistency in tiers of travel, (b) promote sponsored content, or (c) contain outdated information.

In this paper, we present a part of our work which focuses on developing a deeper understanding of travel, and present how such a knowledge can be used to (a) present users with consistent and reliable information, and (b) evaluate machine learning algorithms.

Knowledge extraction tasks based on Wikipedia come in a few flavors: training embedding models [1], building knowledge graphs and ontologies [6], or natural language processing and extraction [3]. Existing knowledge bases like YAGO or DBPedia that build information extraction pipelines using Wikipedia [7] require significant effort to editorially curate the extracted data and label data for training models. Unlike other work in the past, we do not intend to build a knowledge graph or an ontology. Instead, we utilize data from Wikipedia directly. Lots of data are required to build a large ontology or to develop sophisticated machine learned techniques to parse and extract the information from web data. As we shall demonstrate, most of the information can be obtained from Wikipedia using a simple extraction process with almost no natural language parsing. Further, one of the arguments against an unsupervised model is the lack of data, and our approach demonstrates how the representation of data of Wikipedia enabled us to evaluate a machine learning model.

2 KNOWLEDGE IN WIKIPEDIA

In our work, we are interested in extracting specific Wikipedia entities associated with hospitality and travel along with relevant metadata. In particular, we extracted the brand name or the company name, the name of the establishments, their tier of service, and the location. For a hotel, that would mean, we extract the parent company, say 'Wyndham', and then the list of establishments owned by the parent such as 'Days Inn', 'La Quinta', 'Ramada', 'Super 8' and 'Wyndham Grand'. For each of these, we extract their tier of service such as 'upscale', 'mid scale', 'boutique', and 'economy'.

2.1 Information Extraction

The information we seek is predominantly distributed across various organized parts of Wikipedia such as lists, sections, categories, info boxes and templates.

We start with the Wikipedia page: 'List_of_lists_of_lists', which points to several other lists. For the purposes of the work described in this paper, we focused on travel within United States to include list of hotels, airports and car rental companies. In addition to the lists, we also used the Category pages such as 'Category:Vehicle_rental_companies' and 'Category:Travel_and_tourism _templates' as secondary starting points, for topics not listed in the lists. These pages yield the company names and the brand names.

Next, we derive info such as the tier of service and the locations, where appropriate, using the sections, info boxes and subcategories within the pages for each of the brands we extracted.

Finally, we sought human editorial help[1] for verifying various labels such as tiers and categories for the extracted data. Remarkably, the travel activities annotated using Wikipedia extractions

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316759

[1] an in-house editorial team at Verizon Media

agreed with editorial review over 65% of the time. The accuracy was lower than expected since the brand differences are subtle and can fall into multiple tiers of service e.g. upper midscale hotels can be classified as upscale or midscale. This reinforced our belief that data from Wikipedia is sufficiently accurate, and some of gaps in agreement were due to standardization of labels.

2.2 Challenges

While we highlight the ease of information extraction from Wikipedia, it is important to point out the challenges and limitations we encountered.

- **Incomplete Data** While Wikipedia continues to grow daily, the data is not complete. For instance, not all hotels in a city are listed, and lesser known establishments are often missing.
- **Inferring Complex Relationships** Let us consider the example of airports. In the current state, it is not possible to derive information such as nearby airports without advanced processing or potentially building an ontology. While sites like Wikivoyage support complex queries to extract nearby airports, they are limited by the Wikipedia ontology of neighboring attractions to extract nearby airports.
- **Lack of Consistency within Categories and Templates** Across several pages within Wikipedia, the categorization and tiers of service labels are inconsistent and there is some disagreement about the labels, and we have to normalize these labels.
- **Choice of labels for entity categorization** Vehicle rentals could be categorized as car rentals, truck rentals, rideshares etc. or they could be categorized by car design size such as Compact, Mid-size, Family etc. Choosing the appropriate set of Wikipedia labels for categorization is a well-known problem [4].
- **Lack of Consistent formatting within Wikipedia** Not all lists are formatted the same. While some are organized as tables, others are enumerated or are bulleted lists. Parsing such pages takes some effort.

In spite of these shortcomings, we find that the data in Wikipedia is a good starting point for large scale information extraction. Since they cover the most frequent or most popular destinations, commonly referred to as the head, the data is good for initial prototypes.

3 CASE STUDY: UNDERSTANDING TRAVEL SEARCH QUERIES

We studied a time ordered trail of search queries of users from Yahoo! Search logs, with the intent of understanding typical travel related search patterns within user search sessions. The data used was anonymized. Queries such as "cheap flight tickets to Orlando" or "Sheraton hotel address in downtown Atlanta" and "Enterprise rental Dallas airport" are examples of typical queries with travel intent. We built an embedding model to derive the relationships between search queries, and we used data extracted from Wikipedia to evaluate our model. We describe each of the process below.

Table 1: Geographical and Brand-Tier similarity scores for all travel activities

	NN=1	NN=2	NN=3	NN=5	NN=10
$Geographical$	0.536	0.502	0.479	0.446	0.403
$Brand-Tier$	0.552	0.528	0.514	0.498	0.476

Table 2: Geographical and Brand-Tier similarity scores for vehicle rental activities

	NN=1	NN=2	NN=3	NN=5	NN=10
$Geographical$	0.583	0.534	0.506	0.465	0.415
$Brand-Tier$	0.633	0.600	0.569	0.523	0.462

3.1 Query Embedding

We train an embedding model similar to Word2Vec[5], where users are equivalent to a document, search queries are the 'words' and a search session is a 'sentence'. Similar models have been proposed in the past for query understanding [2] to generate a 300-dimensional embedding for each search query. In the resulting embedding space, related travel activities appear closer to each other. We extract the top 100 neighbors of each travel activity using cosine similarity for each travel activity pair. We set the threshold frequency for travel queries to have at least $f = 500$ occurrences for generating the activity embeddings.

3.2 Query Evaluation

For each travel activity query, we compute the similarity scores for the top k nearest neighbors where $k = \{1, 2, 3, 5, 10\}$. For our case study, we computed similarity scores across 2 dimensions:

(1) Geographical similarity
(2) Brand-Tier similarity.

We augment a given travel query using the domain knowledge from Wikipedia sections that were extracted using the process described earlier in Section 2.1. The augmented data is then used to compute the similarity scores. For example, a traveler searching for "cheap flight tickets to Orlando" could also be potentially interested in "budget hotels near Disney Orlando" or "economy car rentals in Orlando airport". We expect such queries to be more related to each other and other queries related to Disney than a query like "hotels near Aspen ski resort".

3.3 Results

Table 1 shows both geographical (city or state) similarity and the brand level or category level similarity scores across measured on a set of 20k activities for the top nearest neighbors. An $NN = 1$ score of 0.536 for the *Geographical* score indicates that 53.6% of the first-nearest neighbors of queries had the same geographical intent (same city). The similarity scores decrease with the increase in the number of neighbors. This is consistent with the fact that the farther neighbors are less relevant than the nearer ones.

As observed, the $Brand-Tier$ scores are better than $Geographical$ similarity scores across all travel activities. One reason for poor $Geographical$ scores is while flights and cars are reserved at specific city airports, the corresponding hotel reservations could be at a nearby city suburb where the intended travel activity is performed e.g. while a flight reservation is made for Dulles airport in Washington DC, the hotel reservation could be made in Alexandria, Virginia. This effect is more pronounced for specific activities like vehicle rentals as shown in Table 2. On the other hand flight reservations exhibit better geographical similarities than other activities.

4 CONCLUSION

In this paper, we presented a part of our work in progress to that demonstrates the use of Wikipedia to develop a deeper understanding of the travel domain. In particular, we described an approach to extract entities and their labels, and showed how to use the extracted information to evaluate an embedding model. Subsequent work is aimed at improving the extractions to increase of the depth of knowledge. We are also exploring ways to solve some of the challenges described in Section 2.2.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR* abs/1607.04606 (2016). arXiv:1607.04606 http://arxiv.org/abs/1607.04606

[2] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. 2015. Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 383–392. https://doi.org/10.1145/2766462.2767709

[3] Lili Kotlerman, Zemer Avital, Ido Dagan, Amnon Lotan, and Ofer Weintraub. 2011. A Support Tool for Deriving Domain Taxonomies from Wikipedia. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Association for Computational Linguistics, 503–508. http://aclweb.org/anthology/R11-1069

[4] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep Learning for Extreme Multi-label Text Classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 115–124. https://doi.org/10.1145/3077136.3080834

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959

[6] Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a Large Scale Taxonomy from Wikipedia. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2 (AAAI'07)*. AAAI Press, 1440–1445. http://dl.acm.org/citation.cfm?id=1619797.1619876

[7] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85. https://doi.org/10.1145/2629489