Spatio-Temporal Quality Issues for Local Search

Dirk Ahlers
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
dirk.ahlers@idi.ntnu.no

ABSTRACT

Geographic search is routinely used in many services and applications that exploit the availability of Web content which is related to a real world place, region or object. However, do you trust the location information? Who has not made the experience that the restaurant you went to has just moved to another part of the city or shut down? Local search returns located results, e.g., extracted entities located in a certain spot or area, but their quality can be difficult to judge. Compared to normal Web search, local Web search has additional inherent issues due to factors such as insufficient semantics, ambiguity of references, imprecise mapping, or unknown status of the real-world entities described in documents. We present selected issues and features of geospatial quality and credibility based on spatial, temporal, and topical indicators as an additional measurement of spatial relevance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Search quality; credibility; location-aware search; geospatial Web retrieval; entity extraction; temporal retrieval

1. INTRODUCTION

Location is an important and successful feature in Web information retrieval. Scenarios include the planning of a business trip or a vacation, the search for a new house, the decision for a school or university or something as simple as where to buy groceries [1, 5]. The reliability of search results becomes even more crucial in a location-based or mobile scenario where it can be much more difficult to judge the reliability of information short of actually visiting a place. Consider a user's dissatisfaction who drove all the way to an out-of-business restaurant that was, however, shown in a list of results of recommended restaurants. The difference in perception to Web search can be explained by the fact that a result document matches the query in some way, but a result entity carries the implied assumption that it is actually represented in the real world – and it is much easier to check a document than it is to check an actual place. In short, the existence of information about a place is expected to be a proxy for the existence of the place itself.

As an example, consider a query about the "Museum of Modern Art" (MoMA). The usual result would be the information about its exhibitions, its homepage (www.moma.org) and of course its location. Normally this would be given as 11 West 53rd Street, Manhattan, New York, NY, USA, or, with less granularity, as New York, USA. However, the query is broader than the example suggests. There is also the San Francisco Museum of Modern Art (SFMOMA) and of course many more museums with this exact name around the world. Optimally, these multiple occurrences are identified, disambiguated, and treated separately.

The issue becomes more interesting when taking more complex situations into account. For example, changes in the real world can leave marks in the Web that can be challenging to interpret. A special event took place between 2002 and 2004. The MoMA underwent renovations and was closed during that time. The museum itself was unavailable to visitors, but parts of the collection were exhibited elsewhere, some in the MoMA QNS which was located at 33rd Street and Queens Boulevard. When further searching for MOMA in Queens, the P.S.1 Contemporary Art Center comes up at 22-25 Jackson Ave, Long Island City, NY, USA. Yet this is is a separate entity only affiliated on some exhibitions with the MOMA. A major part of the collection was further shown in Berlin during 2004 under the name "Das MoMA in Berlin" at the museum Neue Nationalgalerie at Potsdamer Straße 50, 10785 Berlin, Germany, which received a large media response. Nowadays, the exhibition is gone, but many Web pages still mention it and it turns up in local search engines. So in their own right, all of these results would be valid for a certain period of time only.

Figure 1 shows a result set for "museum of modern art" on a global scale, highlighting New York and Berlin; the cutout shows Manhattan and Queens. For these three results, their lifespan is plotted in the graph; the shaded shapes give an estimate on expected pages mentioning the place. Web resources may not always match very well to the entity lifespan and the evidence for a new entity is usually much better than that for its disappearance. While a global search could very well judge the MoMA in New York to be the most important

HT'14, September 1–4, 2014, Santiago, Chile.
ACM 978-1-4503-2954-5/14/09.
http://dx.doi.org/10.1145/2631775.2631792 .

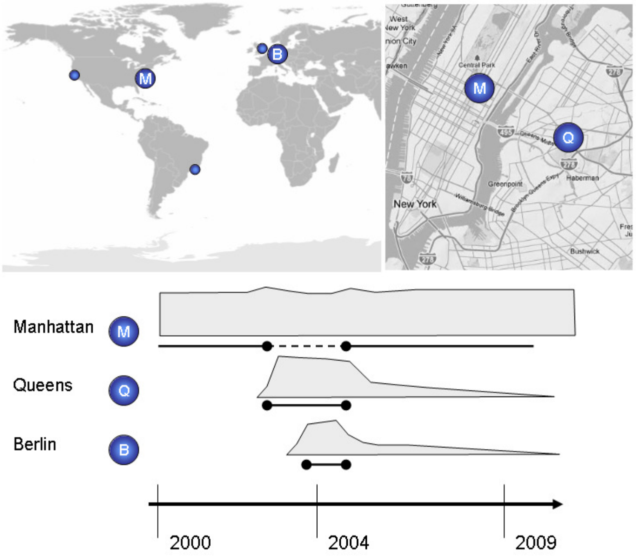

Figure 1: Example of temporal changes for entities

one, there usually would be results all over the map. A local search within Berlin would still find strong evidence for its presence.

This article presents an initial discussion of credibility issues influencing geospatial search quality and relevance.

2. QUALITY AND RELEVANCE

The Web is a major information source, but also contains inaccuracies, omissions, outdated pages, or even adversaries. Thus, retrieval techniques have to address the information credibility issue to deliver high-quality results. Geospatial Web search engines [4] face similar issues, but often on different levels or, due to the spatial processing, from entirely new angles. User requirements include not only relevance, but also features such as freshness and diversity [18] or reliability. In Geographic Web information retrieval [13], results are ranked based on the relevance of the documents to a query on content and spatial features [9, 10, 16]. Our aim is to address location data quality to lay the ground for higher-level quality models. Then, quality and credibility is added as another dimension of relevance:

$$\begin{aligned} Relevance(q, doc) \quad = \quad & Rel_{Content}(q_{content}, doc_{content}) \otimes \\ & Rel_{Spatial}(q_{spatial}, doc_{spatial}) \otimes \\ & Credibility_{Spatiotemporal}(q, doc) \end{aligned}$$

Going into more detail, we understand geospatial credibility not only as individual data quality, but as a new compound measure across a variety of features and sources such as reliability, authority, trustworthiness, quality, correctness, up-to-dateness, completeness, availability, and finally, physical existence in the real world. Note that credibility not only depends on the documents in the index, but also on the query, as different queries can have different expectations or requirements.

2.1 Selected Factors

Credibility indicators can be derived from features such as presence of keywords, references, named entities, detectable patterns, times and dates, linkage patterns, or external sources, to name just a few. However, in many cases even these features are not completely accurate or carry a certain uncertainty [6, 11], as well as their extraction process, either implicit or explicitly annotated. Therefore, the following factors indicate uncertain data on multiple levels.

Temporal features are strongly connected to geospatial information [7]. A question for museums in the vicinity implicitly has a temporal aspect in that the museums should be there at the time of the question. Web pages frequently appear, disappear, move, or change. Similarly, the location references on them and those in the real world do the same when businesses are created, move, change name or owner, or are closed. For the example of the MoMA Berlin, it would usually take some time to show up in search results. More importantly, it is more difficult to detect its closing. An opening of a new place thus is harder to detect than a closing or a move of a business. This also applies to compound or aggregate measures.

Geoparsing and Geocoding issues to consider include extraction errors, wrong matches, and granularity, [12, 17]. In the example, using only broad granularity would identify both the Manhattan and Queens buildings as New York.

Entity and Relation Extraction plays a major role in location credibility within entity-oriented search such as company or people search [8, 17, 3]. Issues arise with incomplete entities due to varying levels of detail or spellings of names, imprecision in extraction etc.

External Data Integration of spatial domain knowledge is necessary for, e.g., validation, but it may also already carry inaccuracies [2] and might change over time.

Aggregation can increase the reliability of results that are based on multiple sources [15, 3, 14]. For example, some pages may show outdated information or report different names for a museum. We would have to identify MoMA with Museum of Modern Art and similar names, yet distinguish them from the MoMA QNS and maybe even from the museum store within the same building, while considering the temporal changes.

Visualization and interaction can draw user's attention to the reliability of information or may compute aggregate mappings to achieve its goals [14]. This is especially important as items on a map tend to be trusted more by users since mapped data "feels" more accurate.

3. CONCLUSION

We have discussed quality and credibility issues in geospatial Web information retrieval on an initial selection of spatio-temporal features. Ultimately, they should be incorporated into an improved model for relevance ranking. Temporal issues are only one aspect of a more complete model of credibility, but they show up in many features and are underlying non-considered causes for certain cases of lowered result quality. Future work will investigate cross-relations and propose solutions or mitigations for certain issues as well as connect these issues with requirements and implications for different application scenarios. We believe that these issues will have to play a bigger role in future geospatial retrieval systems and open up interesting research questions towards improved trust and quality.

4. REFERENCES

[1] D. Ahlers. *Geographically Focused Web Information Retrieval*. OlWIR, Oldenburg, Germany, 2011. PhD Thesis.

[2] D. Ahlers. Assessment of the Accuracy of GeoNames Gazetteer Data. In *GIR '13*, 2013.

[3] D. Ahlers. Business Entity Retrieval and Data Provision for Yellow Pages by Local Search. In *Integrating IR technologies for Professional Search Workshop @ ECIR2013*, 2013.

[4] D. Ahlers. Towards a development process for geospatial information retrieval and search. In *WWW '13*. 2013.

[5] D. Ahlers and S. Boll. Location-based Web search. In A. Scharl and K. Tochterman, editors, *The Geospatial Web*. Springer, 2007.

[6] I. Askira Gelman and A. L. Barletta. A "quick and dirty" website data quality indicator. In *WICOW '08: Proceeding of the 2nd ACM Workshop on Information Credibility on the Web*, 2008.

[7] K. Balog and K. Nørvåg. On the use of semantic knowledge bases for temporally-aware entity retrieval. In *ESAIR '12*, 2012.

[8] M. J. Cafarella, J. Madhavan, and A. Halevy. Web-Scale Extraction of Structured Data. *SIGMOD Rec.*, 37(4):55–61, 2008.

[9] P. D. Clough, H. Joho, and R. Purves. Judging the Spatial Relevance of Documents for GIR. In *ECIR'06*, 2006.

[10] P. Ehlen, R. Zajac, and K. B. Rao. Location and Relevance. In *LocWeb '09*, 2009.

[11] B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and M. Treinen. What Makes Web Sites Credible?: A Report on a Large Quantitative Study. In *CHI '01*, 2001.

[12] D. W. Goldberg, J. P. Wilson, and C. A. Knoblock. From Text to Geographic Coordinates: The Current State of Geocoding. *URISA Journal*, 19(1):33–46, 2007.

[13] C. B. Jones and R. S. Purves. Geographical Information Retrieval. *International Journal of Geographical Information Science*, 22(3):219 – 228, 2008.

[14] C. Kumar, W. Heuten, and S. Boll. A Visual Interactive System for Spatial Querying and Ranking of Geographic Regions. i-KNOW '13, 2013.

[15] R. Lee, D. Kitayama, and K. Sumiya. Web-based Evidence Excavation to Explore the Authenticity of Local Events. In *WICOW '08*, 2008.

[16] B. Martins, M. J. Silva, and L. Andrade. Indexing and Ranking in Geo-IR Systems. In *GIR '05*, 2005.

[17] Y. Morimoto, M. Aono, M. E. Houle, and K. S. McCurley. Extracting Spatial Knowledge from the Web. In *SAINT '03*. 2003.

[18] I. Szpektor, Y. Maarek, and D. Pelleg. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. WWW '13, 2013.