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 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.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author. Copyright is held by the owner/author(s).

ACM 978-1-4503-2954-5/14/09.
http://dx.doi.org/10.1145/2631775.2631792
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:

\[
\text{Relevance}(q, \text{doc}) = \text{Rel}_{\text{Content}}(q_{\text{content}}, \text{doc}_{\text{content}}) \otimes \text{Rel}_{\text{Spatial}}(q_{\text{spatial}}, \text{doc}_{\text{spatial}}) \otimes \text{Credibility}_{\text{spatia}temporal}(q, \text{doc})
\]

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. 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.

2.1 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


