A Web more Geospatial: Insights into the Location Inside

Susanne Boll
University of Oldenburg
Germany
susanne.boll@uni-oldenburg.de

Dirk Ahlers
OFFIS Institute for Information Technology
Oldenburg, Germany
ahlers@offis.de

ABSTRACT

The Web today is considered to be a sheer unlimited resource of interlinked information which can be explored following links or can be found employing keyword-based search engines. A feature that becomes more and more relevant for our search and use of the Web is the geospatial reference of information. In this paper, we understand the Web as a vast geospatial information space in which most of the location information is still hidden inside the Web's content. We discuss the processes of uncovering hidden spatial information on the Web to realize a multitude of geospatial user scenarios. To explore the spatial character of the Web, location information needs to be discovered, understood, and augmented. By providing location insights into the existing Web, its content becomes accessible to spatial applications and thus allow users exploring the geospatial Web.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H 5.4 [**Information Storage and Retrieval**]: Hypertext/Hypermedia.

General Terms

Design, Experimentation, Human Factors, Management, Measurement

Keywords

Location, geospatial Web, location semantics, location-aware Web search, geographic Web information retrieval

1. INTRODUCTION

Different applications such as car navigation systems, Web map services, and mobile search made location-based services and applications popular for end consumers. We believe that we are just at the beginning of a "geo wave" in which location will become relevant for a multitude of Web applications. Location will be a major driving force behind future Web development. In our view, geospatial information need not to be "added" it is already there and waits to be uncovered and used in a future Web.

To fully appreciate the wealth of information the Web has to offer, we need to understand the role of location within the Web, its structure, content, metadata. We also need to look into the relevance of location-information for the users of the Web. What kind of spatial information do users want? What solutions can best meet to their information needs? How can we satisfy this need? And how can we use and improve the Web in the process? The claim we make is that we do not have to wait for a next generation of the Web that is location-aware but that we have to develop technology to uncover spatial knowledge from the Web and enable users to search, visualize, and explore the Web content with respect to its location semantics. For this, we have to understand the spatial character of the Web to find, understand, augment, and explore these location semantics and unfold as-yet non-geo-referenced Web content in a spatial Web.

In the following we will continue this motivation for understanding the hidden location-information in the Web in a scenario. We present the processes needed to discover, understand, augment and explore the Web with regard to its location semantics.

Copyright is held by the author/owner(s).
Webevolve2008 at WWW2008, April 22, 2008, Beijing, China.

2. SCENARIO

Today, Web search is the method of choice for users to access Web content. Regarding the role of search engines as gatekeepers, some would say "You do not exist if you cannot be found". From our location-based point of view, we would extend this and say, "you do not exist in a spatial Web if you cannot be located".

The spatial character of the Web and its content can be the basis for innovative approaches in structure, retrieval and search, both commercially and in research. The current Web interaction paradigm of search, get result-list, click on first link, would—for geospatial tasks—then change into a much more targeted and spatially visualized exploration of the location-based results.

Imagine a Web where geospatial information can be naturally gathered, combined, and integrated into content, structure, and services. With this location insight, you would be able to retrieve in-depth geospatial information about the page you are currently visiting, its location, its community rating, spatially neighboring pages, its geographic audience distribution, how far it is away from your own Web page and much more. Such a geospatial Web invites you to browse and explore the spatial connections: Planning your next vacation, you find a set of nice hotels by the sea. Musing that you would like to go diving once again you explore all diving areas in their vicinity to choose the best place for the vacation. Enjoying a nice vacation, you feel you have to have a new scuba gear. You search shops and places around and find that a shop close by your hotel offers just the brand you

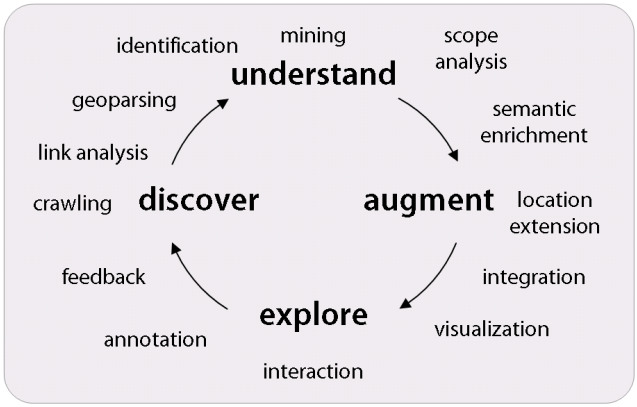

Figure 1: Processes for uncovering hidden geospatial information on the Web

prefer. The search also leads you to interesting local stories about the old harbor where the store is located in and a Wikipedia article about it.

This scenario reminds us of Vannevar Bush's vision of the Memex [6] which has served as inspiration for today's hypertext systems. But if we look closer, we find that the Memex did not only have preconceived connections, but would also drive on the "trails" of a user navigating through it and finding, verifying, and establishing connections on his own which would only then become evident in the system. We feel that this is a very good metaphor for the potential of a geospatial Web. Not all possible meanings are already captured and not all possible connections are yet drawn. Using geospatial reasoning, we could understand geospatial properties of Web content better and based on this understanding, offer further connections between pages and media that can open a new perspective onto the Web. A user will be able to find a georeferenced Web page, interact with its spatial relations, and explore the geospatial dimension of the Web to gain new insights into existing information.

Parts of the described scenario are addressed by individual services or data sources today. Still, the user currently has to painstakingly collect and combine this information on her own without powerful tools to aid in analysis and visualization within a integrated view on all data. The fundamental idea is to automatically create and annotate these connections, to allow geospatial navigation on the existing Web content, and to offer tailored navigation experiences to the users.

In the following, we present the central challenges we see on the way to such a location-aware Web. Figure 1 summarizes the four main processes and tasks we identified for uncovering and exploring hidden spatial information in the Web. The remainder of this paper is structured according to these processes.

3. DISCOVER

Web search engines have been evolving along with the Web for a long time. Just as we need textual search engines to ask for the textual content of documents, we now need location search engines to ask for location semantics. Yet, a simple request for an authoritative page for a given region or the number of Web pages in a 5km radius around a conference venue proves to be most difficult in answering and requires new approaches in resource discovery, information retrieval, and location understanding.

Supporting the high significance of location-based information for users, a study in 2004 [15] concludes that up to 20% of user queries express a geographic information need. A more recent study on search behaviour on mobile devices [11] gives an estimate of 5–15% for location-related queries. We see a significant amount of Web pages today already containing viable location information, but simply not in a semantically structured way. According to "experiments with a fairly large partial Web crawl", [14] found that "approximately 4.5% of all Web pages contain a recognizable US zip code, 8.5% contain a recognizable phone number, and 9.5% contain at least one of these". The full range of location information includes such simple keywords, a brief mention of a region or place or a precise reference to a specific place. The relation of Web content to a physical location is hardly semantically captured in the page but rather implicitly part of the content as an address or a place name [3]. Even though these location-related pages represent only a fraction of all Web pages, their relation to a location is a precious and yet unused asset for exploration in location-aware applications. The field of geographic Web information retrieval aims to unlock implicitly hidden geo-references within Web resources to allow understanding the spatial properties of individual documents and following, the spatial character of the Web.

Today, several standards for description and exchange of location data on the Web exist with various power of expression. These range from simple coordinate specifications for latitude and longitude such as metatags, vCard, Microformats or W3C Geo to more powerful formats able to express additional concepts like lines, boxes, polygons etc. such as Dublin Core Metadata, GML, KML or GeoRSS. The description of a location is typically accomplished in either of two ways: specification of a coordinate tuple, or more often a named hierarchical description. However, only few Web pages describe location information explicitly with any of these standards and most of the location information to be found is simply written into the content of a document. We can conclude that with Semantic Web Technology [5] we can expect to gain and explore structured spatial knowledge, but will have to wait for the coming years for a wider spread.

The little geo-referenced information on the Web is mostly manually annotated with the above mentioned standards or tagged. The most prominent location-related information are yellow pages. Content from such directories can be searched and is supported by map-based services ("local search"), e.g., Yahoo! Local or Google Local Search. Also Web 2.0 sites such as Mappr, Flickr, or Placeopedia allow to associate content with a coordinate on a map and thus manually geo-code it. Additionally, certain Web directories such as dmoz.org organize Web links according to geographical classification. Hierarchies of places and place names are provided by gazetteers [10], for instance the Getty Thesaurus of Geographic Names. However, this spatial information covers only a very small fraction of all Web content leaving the majority of usable location information within common, unstructured Web pages yet to be discovered.

Many authors express a rather gloomy outlook on the thorough integration of structured, high-level metadata annotation. While we would embrace detailed semantic loca-

tion annotations on the Web, we currently can only concur and treat it as the exception rather than the norm.

We therefore expect that a gespatial Web can and will be built based on existing location information in the Web. The challenge for this is to actually to retrieve spatially related resources. Since there exist not reliable structural hints as to which resources contain relevant location information, a resource discovery process must be employed to gather all relevant pages. One way is a resource-intensive broad search of geospatial Web pages, another is to focus the location insight by topical resource discovery. Focused crawling as introduced by [7] is a crawling strategy based on an assumed cohesion within the Web's link graph. It is a trade-off approach to weight completeness against limited resources. Concerning the applicability of focused crawling to the geospatial domain, we could show in [2] that it can deliver a higher efficiency that common crawling, thus delivering evidence that the geospatial topic weakly corresponds to the Web link structure. Further research into link structure and location semantics will bring valuable insights into this topic.

4. UNDERSTAND

Once location-bearing Web content is discovered, its spatial semantics need to be extracted and understood to enable location-aware applications. Unstructured content has to be analyzed for location relations and spatial semantics. The issue of ambiguity could be solved much better with semantic annotations. Without semantic annotations to tell a machine exactly what content creators had in mind, we have to rely on sophisticated heuristics and algorithms to guess at their intention with regards to content and location. Knowledge and intention are inherently hard to capture, understand, and model. However, since we are getting increasingly better at automated content analysis, we can infer some of the intentions automatically even on non-annotated unstructured content. One drawback is that we might be forced to add a further meta-description, namely the accuracy of our "guess".

A vital prerequisite is the reliable identification, disambiguation, and verification of geographical entities [14], especially considering the ambiguities and uncertainties characteristic to the geographical context. Further steps during geographical processing include the detection and identification of geographical context, subsequent processing and data mining; the use of metadata and ontologies in assigning geographical scope and concepts.

Once information has been derived by thorough content and context analysis, it can then be annotated and augmented, be analyzed, fed back, published, and be used for further analysis. Such structured information can be used to understand the role of the location information with regards to the resource and augment it accordingly.

5. AUGMENT

Understanding the location of Web content initially results in any form of location information. For a profound and rich understanding of the spatial character of the Web, however, we need to augment and integrate this location information with other geospatially related knowledge sources. Hence, geospatial content can be semantically enriched by combining diverse sources of information which can be based on low-level analysis, associated contextual information as well as domain knowledge. Further augmentation can be provided by the extraction of specific content or context features and also the intelligent combination of metadata to derive higher-level semantics. Specifically targeting at geospatial search and retrieval, an interesting bearer of location-information can be Web images embedded within a page. The connection of textual and media content can be explored [1] with regard to a shared location. Similar explorations of relations by Web structure will draw on the focused crawling approach detailed earlier.

As a very valuable source for semantic enrichment we see external sources from the Web like for example photo community sites, Web gazetteers or online encyclopedias which are also at least in part geo-referenced. Approaches in Internet Mapping try to acquire and save topological data of the Internet infrastructure [12] and create maps on a technical network level. This information could also be integrated into a holistic location model for the Web. Hierarchies of place names are often used for the identification and parsing of geographical information. Such domain knowledge of spatial information can be found in taxonomies, gazetteers, thesauri, and geographical ontologies. This knowledge can also be used for disambiguation or for semantic enrichment of Web content by accessing and matching additional information which is not implicitly part of the location description. Drawing on this source, previously unknown connections can be uncovered.

We also see interesting work on extending a location relation present on a page to a more thorough understanding of the page's geographical context. These approaches include the definition of a target audience, an actual audience [4], relationship to similar pages, geographical footprint [8] etc. The semantics of a location on a page can thus be manifold. Also the location information might apply to Web content in different granularity from an entire page Web site down to small fraction of a Web page.

When we search the Web we assume that the retrieved results are attempts to best meet the user's query. The reliability, however, of these results become even more crucial in a location-based scenario. Consider a user's dissatisfaction who drove all the way to an out-of-business restaurant that was, however, shown in the list of results of restaurants nearby. There will always remain a level of uncertainty, but we need to develop a transparent model of trust and up-to-dateness that allows a user to better understand and value the results of their spatial search.

Augmentation deepens and specifies the spatial context of Web content and contributes to better spatial experience of Web content explored by it's users. The spatial knowledge extracted, analyzed, and augmented in different processes today resides in providers, applications, and early research prototypes. In the future, we expect this knowledge to become part of and available in a spatial Web.

6. EXPLORE

The search for spatially related information is becoming widespread. Technologies offered to users today to efficiently search for this information range from keyword-based searches on prepared geospatial data sets to map-based local searches or natural language queries for spatially related content. In the field of data visualization and exploration we see the ongoing development of interaction models that

allow to understand the spatial semantics of data sets and to derive further spatial knowledge from the user interaction. Task-oriented visualization and interaction models help to understand, mine and augment spatial relationships of Web content.

Visualization of data with geographical relations has specific demands which has lead to an own research area known as Geovisualization [9]. In contrast to the traditional field of cartography—the creation of maps—geovisualization is characterized by a rather strong interactive component for both the presented content and the presentation itself. Geovisualization has reached the end user on a large scale with applications such as Google Earth and NASA World Wind in which overlays allow to add and visualize information for any place on the map. If this can be combined with geospatial Web information we arrive at geographical mashups in which freely-available functionality and data are loosely but flexibly combined.

To enable users to access the spatial character of the Web, human spatial cognition can be the key to further the development of user interfaces. The main insight we can expect from this field is that *presenting the structure of an environment helps people to understand the environment*, i.e., the spatial cohesion and spatial relations of the Web. This further motivates the demand for tailored visualizations of the geospatial Web. Learning from users and also allowing them to heavily interact can be used to provide feedback to improve all other areas discussed here, by giving hints towards new resources, aiding, guiding and rating an automatic retrieval process and interaction within the visualization.

7. CONCLUSION

We have presented challenges and ideas that mark the way towards delivering location insights into the geospatial dimension of the existing Web to the user. Only when the described processes are integrated and their respective challenges are met will we truly be able to activate geospatial information as a further natural dimension of Web search and interaction. We have to cope with the fact that Web pages are designed to be read by people instead of machines. But we should also keep in mind that Web pages are designed by people and not machines.

The emergence of microformats and similar techniques as a way to integrate semantics into common Web pages shows that there is a desire for more structured annotation, but in an easy, light-weight way, similar to tagging [13] as a powerful organizational metaphor. This currently seems to be all we can hope for on a large scale. Still, unstructured location information is created daily on the Web and is an opportunity too good to miss. What is needed is not only more location annotation, but first of all more intelligent analysis that exploits what is already in our hands. We must be able to understand location-relevant content so well that we can annotate it automatically and augment and connect it to further location-related knowledge. To finally convey the spatial character of the Web to the end user, we need suitable visualization and interaction techniques that show and allow exploring the content, relations, and connections of a geospatial Web.

What will the future geospatial Web look like? We cannot be entirely sure, but we expect it to be a combination of existing standards and existing content with new analysis and retrieval techniques based on improved models that are adapted to the spatial character of unstructured Web content. Accessible by suitable visualization interfaces, users can then intuitively experience the vast geospatial information space the Web has to offer. Location information is an existing but yet mostly unused asset that forms a great opportunity in a geospatial Web.

8. REFERENCES

[1] D. Ahlers and S. Boll. Oh Web Image, Where Art Thou? In S. Satoh, F. Nack, and M. Etoh, editors, *Multimedia Modeling 2008*, volume 4903/2008 of *LNCS*, pages 101–112, Kyoto, Japan, 2008. Springer.

[2] D. Ahlers and S. Boll. Urban Web Crawling. In *Location and the Web (LocWeb2008) Workshop held at WWW 2008*, Beijing, China, 2008.

[3] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web Content. In *SIGIR '04*, pages 273–280, New York, NY, USA, 2004. ACM.

[4] S. Asadi, J. Xu, Y. Shi, J. Diederich, and X. Zhou. Calculation of Target Locations for Web Resources. In *WISE*, pages 277–288, 2006.

[5] T. Berners-Lee and M. Fischetti. *Weaving the Web*. Harper San Francisco, September 1999.

[6] V. Bush. As We May Think. *The Atlantic Monthly*, 176(1):101–108, July 1945.

[7] S. Chakrabarti, M. van den Berg, and B. Dom. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. *Computer Networks*, 31(11-16):1623–1640, 1999.

[8] J. Ding, L. Gravano, and N. Shivakumar. Computing Geographical Scopes of Web Resources. In *VLDB 2000*, Cairo, Egypt, 2000.

[9] J. Dykes, A. MacEachren, and M.-J. Kraak. *Exploring Geovisualization*. Pergamon, 2005.

[10] L. L. Hill. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *ECDL 2000*, pages 280–290, London, UK, 2000. Springer.

[11] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *CHI '06*, pages 701–709. ACM, 2006.

[12] A. Lakhina, J. W. Byers, M. Crovella, and I. Matta. On the Geographic Location of Internet Resources. In *IMW '02*, pages 249–250. ACM, 2002.

[13] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. In *HYPERTEXT '06*. ACM, 2006.

[14] K. S. McCurley. Geospatial Mapping and Navigation of the Web. In *WWW '01*, pages 221–229, New York, NY, USA, 2001. ACM.

[15] M. Sanderson and J. Kohler. Analyzing geographic queries. In *ACM SIGIR Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004.