Project 3: Linking the Virtual and Physical World

Virtual location-based services (VLBS) aim to link physical locations with virtual locations (e.g., a website). Within our prototype, we currently exploit existing connections between virtual and physical locations derived from the Google Places knowledge base. Here, many places such as hotels, shops, restaurant, companies and other are assigned with an URL, typically the website of the hotel shop, restaurant, etc. In the long run, it is conceivable to connect all kinds of locations.

Linking virtual locations. Consider news articles on different news sites that cover the same story. Intuitively, users reading an article about Obama ordering air strikes but either on CNN or BBC can be considered to be at similar virtual locations. However, linking those virtual locations is challenging for multiple reasons:

- Measuring the similarity between virtual locations. Simply speaking, virtual locations are Web pages or sites, and quantifying the similarity is non-trivial. For most applications, two pages are similar if they address the same topic. Measuring the similarity between documents using, for example, the Vector Space Model, is a common task, but assumes clean documents. On Web pages, however, actual content is hidden in HTML markup and among content-unspecific elements like advertisement, navigation menus or links to other content. Extracting the main content of a page is a research topic by itself. The result of such an effort is Boilerpipe\(^1\), a library providing algorithms to detect and remove the surplus clutter around the main textual content of a Web page.

- Extending users' presence across multiple virtual locations. Using a straightforward approach, a user can be at the virtual locations reflecting the Web page the user is visiting and all virtual locations with a similarity above a certain threshold. This solution is rather easy to implement (once the pairwise calculation of similarity between virtual locations is established). However, as a result, the areas of presence of different users can overlap. An alternative approach is to group virtual locations based on their similarity into distinct groups. From an algorithmic perspective, this task of grouping translates into a clustering problem with an unknown number of clusters.

Linking Physical and Virtual Locations. Exploiting Google Places to find physical locations that are assigned to a URL is intuitive and easy to implement but represent only a first (and minor) step to elicit the full potential of connecting the Web with the real world. For example, a hotel is not only represented by its website on the Web, but also by reviews on recommender sites or personal blogs. Whether a user is visiting the hotel's website or reading a review about the hotel on Tripadvisor, the user's context of interest is the hotel.

For selected websites such as recommender sites that feature detailed and structured information about places, automatic approaches to link a review to the corresponding physical location might be applicable. Here, website-specific crawler can extract any kind of available information about the location, e.g., postal address, and perform an address lookup to find the location's geo coordinate. Such crawler are typically very easy to implement (less than 1 day's work) and a meaningful method for connecting physical with virtual locations in an automatic manner and with a high accuracy. To extract real-world places from arbitrary Web pages is much more complex. NER might be applicable, but without a (very) accuracy this might lead to a rather "dirty" link dataset.

---

1 https://code.google.com/p/boilerpipe/