A Map-Based Web Search Interface Using Point of Interest Aggregation

Kwangsoon Jung[1], Sangchul Ahn[1], and Heedong Ko[1,2]

[1] Department of HCI and Robotics, University of Science and Technology
Daejeon, South Korea
[2] Imaging Media Research Center, Korea Institute of Science and Technology
Seoul, South Korea
{ksjung85,prime,ko}@imrc.kist.re.kr

Abstract. With advent of a mobile computing, the pattern of information search has been changed. Search queries through mobile devices increase; 30% of Google's organic search queries come from mobile devices, and local search, which seeks information with geospatial constraints, also increases. As of 2013 local search on mobile phones continues to grow up to 60% since 2010. However, a large number of web documents cannot be exposed to local search even though they refer to a point of interest (POI) just because they are not explicitly geo-tagged. We are interested in connecting typical web documents to spatial search based on POIs by geotagging web documents. In this paper, we present a map-based web search system that serves geospatial search queries for non-geotagged documents. The proposed system provides with fine-grained local search for typical web pages mentioning several POIs and supports semantic search in accordance with their spatial relation of inclusion.

Keywords: Point of interest, map search interface, named-entity recognition, toponym resolution, entity linking, local search.

1 Introduction

With the recent advent of mobile computing, more people seek local information near them. According to the survey [13], as of 2013 the local search, which seeks information with geospatial constraints, on mobile phones continues to grow up to 60% since 2010. The local index storing local information is needed for the users who seek local information. If the user searches the local information with the usual search engine, the usual search engine can give the information mixed local-related information and non-local-related information. For example, when the user queries "Washington" for the purpose of searching local information, the search engine will give mixed results about the person "Washington" and the location "Washington". We wish to build a map interface using the index storing local information.

However, a large amount of indexed documents in the web have no geo-tags, so they cannot be indexed in the index of geographic locations. Therefore they cannot be exposed to local search even if the web documents are talking about geographic locations.

S. Yamamoto (Ed.): HIMI 2014, Part I, LNCS 8521, pp. 342–351, 2014.

For instance, if "Eiffel Tower" is described in someone's blog, the document does not appear in the search result of Paris of geographic constraints on the map without explicit geo-tagging.

Our system is a map search interface system showing documents of no geotags for local search. Our system consists of four parts: DPOI, crawler, poi indexer and map interface. DPOI provides with the POI list to geotag. The crawler crawls the documents in the web. POI indexer takes the crawled document and detects the POIs in the document by resolving the ambiguity problem of multiple locations with the same name and constructs the POI index. Finally, the map shows the geotagged documents on the map by using the POI index. Our strong point is POI. POI is a place representation like restaurant, hotel, store, cities, etc. on a digital map. Existing works[1,2,3,4,9] in this area generally focused on building the geotagging index with cities, states, or countries that are the static geographic locations. Instead, we extend to all types of POIs including restaurant, stores, landmarks, etc. where are relatively dynamic compared to cities, states, and countries. Moreover, the existing works assume that the author writes a document in concentrating on the one geographic region[1,2,7]. Instead of representing one region to a document, we represent the focused region with hierarchical representation, we called POI level of detail (LOD). For example, we can represent the two LODs, U.S.-California-Anaheim-Disneyland and U.S.-Nevada-Las Vegas for a weblog that the author stayed the Disneyland, CA in the morning and moved to Las Vegas, NV in the afternoon.

2 Related Work

Many geo-tagging systems use a gazetteer, which is the database of geographic locations containing latitude, longitude, population, and more, to collect a set of geographic locations and its attributes. Web-a-where[1] constructs a gazetteer from the dataset of GNIS, World-gazetteer.com, UNSD, and ISO 3166-1. Geographic locations have a hierarchical structure of a world divided into continents, countries, states, and cities. STEWARD[2] extends the number of geographic locations from the number of location in Web-a-where. [3] consists the gazetteer of the several category taxonomy by using the place type, geographic containers, and the population of the place. Unfortunately, their common limitation is that they only use the source of the cities, states, countries, and continents, which do not include all types of POIs.

Geotagging involves two problems; toponym recognition and toponym resolution. Toponym recognition is the problem dealing with geo/non-geo ambiguity, which finds geospatial names in the documents. The most common way to find geospatial names is to use string-matching algorithm from a set of geospatial names in a gazetteer. However, the string-matching algorithm alone can lead to the problem. For example, "Of", a city in Turkey can be detected "of", which is a preposition in English grammar[1]. To solve this problem, many papers[1,2] use the POS(Part-of-Speech) to find noun phrase since geospatial names are always in noun form. For a fine detection, Named-Entity Recognition(NER) technique is used[2,7] as well, which classifies noun into person, organization, and location. Detected as a location are more likely to be geospatial names.

Toponym resolution is the problem coping with geo/geo ambiguity, which is the procedure that determines a specific geographic name with multiple locations of the same name. For instance, “Boston” is located in both United States and United Kingdom. When detected “Boston”, we should clarify which “Boston” the author is talking about. One of popular methods is to use the assumption that the author focuses on one area theme while writing a document, so there is one geographic focus in a document. Lieberman et al. decides the one region theme by using the document specific heuristic rules such as dateline of news articles. Web-a-where assigns the scores to each region, continents, countries, states and cities and decides one geographic focus, the highest score region. NewsStand[4] uses heuristic rules depending on document type, in news, a geographic location in a title has more weights. Adelfio et al. uses a Bayesian likelihood model to determine the place category in the data set of table-based place names.

In Information Retrieval, there is a similar problem called Entity Linking. Entity Linking is a method to detect an entity in the document, and then they connect the entity to the knowledge base such as Wikipedia. Entity Linking also needs an entity resolution that, given a mention, which entity is talking about. For example, “Jordan” in the document can be a basketball player, “Michael Jordan” or a professor “Michael Jordan”. One of the known methods is to use Graph-based method[8,14]. They used “random walk with restart”[10] for disambiguating entities with the same name. We can apply this method to resolve the place name.

Map-based interface is used for visualizing geographic semantics on the map by using tags in Flickr[16,17]. Hiramatsu et al. implemented the map-based interface system showing the local POIs on the map.

3 Architecture

3.1 Digital POI (DPOI)

DPOI is a gazetteer collecting the POI entities and their properties including an address, latitude, longitude, review, etc. An initial set of a POI is predefined by our POI aggregation. A POI can have various names according to different languages. We aggregated the multilingual POI titles from Korea Travel Organization(KTO) and Geonames. The credible user can add the additional POI names if necessary. One of the functions of DPOI is to assign the unique URI to a POI entity. This URI is used as the unique tagging id in the document.

3.2 Crawler

In the crawler module, the crawler crawls the documents in the web by following the URLs embedded in the crawled documents. In order to provide users with the useful information, the quality control of the documents is necessary. We crawled the

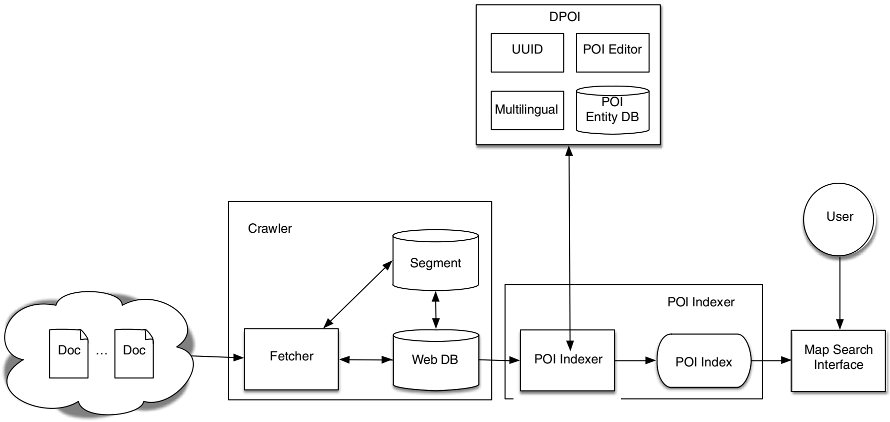


Fig. 1. System Architecture

documents of the reputed users expertising in a trip from Daum-Blog[1] and Naver-Blog[2], Korean famous blog sites. The reputed users are recommended from Daum-Blog, Naver-Blog explicitly by blog-themes for readers. After the crawler crawls the documents in the web, the crawler sends the crawled documents to POI named-entity indexer to detect POIs.

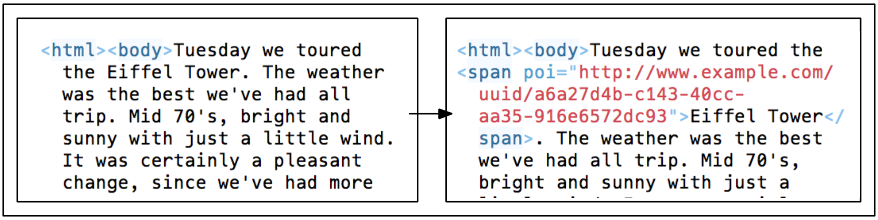
3.3 POI Indexer

When passed the crawled documents, the POI indexer geotags a POI URI to the documents. After the tagging process, the POI index is updated, which is used for search on the map. However, some POI would have a same name, but have different locations. For instance, Boston is both in United States and in United Kingdom. This requires a resolution algorithm, called POI Resolution, to determine which Boston is talking about in the document. We applied a graph-based method[8,14] that a POI mention in the document is likely to be in the same area of a country, state, or city of the other POI mentions in the document. The reason why we choose graph-based method is that if the machine learning technique is applied whenever a set of a POI changes the training should be done again.

3.4 Map Interface

The map interface consists of map, search input box, and search button. When clicks the search button with the keyword, the map interface sends the query containing the map's bounding box, which is the latitude and longitude of left-top, right-bottom of the map, and the keyword to the POI index interface. The POI index finds the POIs

[1] http://blog.daum.net
[2] http://blog.naver.com



(a) (b)

Fig. 2. Two documents before and after geotagging. (a) shows a crawled web document mentioning "Eiffel Tower" POI. (b) shows the document geotagging "Eiffel Tower" as URI in DPOI by POI Indexer.

matched with the keywords in the given bounding box. When clicks the POI, the map interface shows the documents list related to POI. We also provide with the list of POIs on the map for a document. After using our POI Index, the documents without explicit geospatial information can be shown on the Google Map.

4 Graph-Based POI Resolution

We uses Graph-Based POI resolution algorithm[8,14] to resolve which POI the author is talking about. First, we need to construct the POI referent graph. Second, we calculate the initial evidence, which is reinforced by propagating them through the POI reference graph. Finally, the method chooses the highest evidence score, which is the highest probability.

4.1 The POI Referent Graph

The graph is the weighted graph $G(V,E)$ where the node set V comprises of all POI mentions in a document and the POI candidate entities of the mentions in a document. The edge E consists of address edge between POI candidate entity nodes and compatible edge between the POI mention node and POI candidate entity nodes.

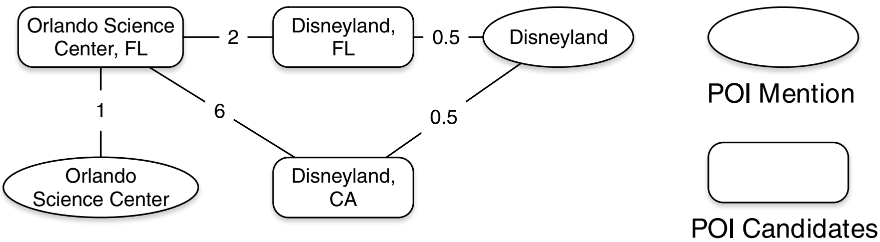


Fig. 3. Example of The Referent POI Graph

Compatible Edge: A compatible edge is an edge between the POI mention node, m, and the POI candidate entity node, e. The total number of POI candidate entities to a POI mention determines the edge weight of the POI. For example, if "Disneyland" has two POI candidate entities, where are in Florida and in California. The edge weights are both 0.5.

$$\mathrm{CP}(\mathrm{m},\mathrm{e}) = \frac{me}{|m||e|} \tag{1}$$

Address Edge: An address edge is a edge between the POI candidate entity nodes. The weight is the address distance[15], which shows the geographical distance of POIs. In order to know the address distance, we need a LOD tree. The LOD tree has nodes that the leaf nodes represent the POI candidate entity and its parents nodes show its county, city, states, countries depending on the depths in a graph. The address distance between leaf nodes is the number of the steps between them to reach in shortest path. For example, the distance between "Orlando Science Center" in Florida and "Disneyland" in California is six. The distance between "Orlando Science Center" and "Disneyland" in Florida is two.

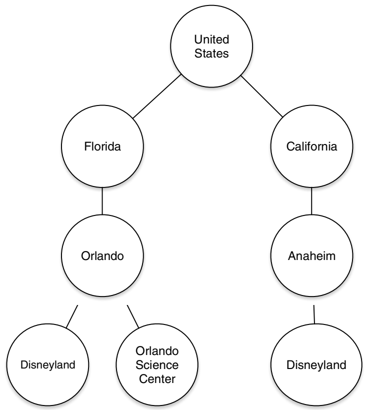


Fig. 4. Example of POI LOD

4.2 Initial Evidence and Evidence Propagation Ratio

The weight of the initial evidence is the score that will be propagating to other nodes. The initial evidence uses tf-idf term score.

$$\mathrm{importance}(\mathrm{m}) = \frac{tfidf(m)}{\sum_{m \in D} tfidf(m)}. \tag{2}$$

For example, if "Disneyland" occurs 8 times and "Orlando Science Center" occurs 4 times, the importance of "Disneyland", importance("Disneyland"), is 8/12.

Given Initial evidence, the next step is to calculate the propagate ratio from the POI referent graph. The propagate ratio is the ratio propagating the initial value to other nodes. The propagating rate depends on the two edges. The propagating rate of Compatible Edge, which is an edge between a POI mention and a POI candidate entity, is below. This is the same score as a Compatible Edge score.

$$\mathrm{P}(m \to e) = \frac{CP(m,e)}{\sum_{e \in N_m} CP(m,e)}. \quad (3)$$

The propagating rate of Address Edge, which is an edge between POI candidate entities, is below.

$$\mathrm{P}(e_i \to e_j) = \frac{AD(e_i,e_j)}{\sum_{e \in N_{e_i}} AD(e_i,e)}. \quad (4)$$

For example, in Figure 3, the propagating rate of the address edge from “Orlando Science Center, FL” to “Disneyland, FL” is 2/6.

4.3 Collective Inference

Given a POI mention m in a document d, we decide the right POI entity as:

$$\mathrm{m.e} = \mathrm{argmax}_e\, CP(m,e) \times r_d(e). \quad (5)$$

where $CP(m,e)$ is the edge weight between a POI mention m and its POI candidate entity e and $r_d(e)$ is the evidence score for each document d and POI candidate entity e. We need to calculate $r_d(e)$. Given the referent graph $G(V,E)$ contains n nodes (n = the number of POI mentions + the number of POI candidate entities). Each node is assigned an integer index from 1 to $\|V\|$. We construct the referent graph G as a matrix where $G[i,j]$ is the edge weight between node i and j

$$r^0 = s;\ r^{t+1} = (1-\lambda) \times T \times r^t + \lambda \times s. \quad (6)$$

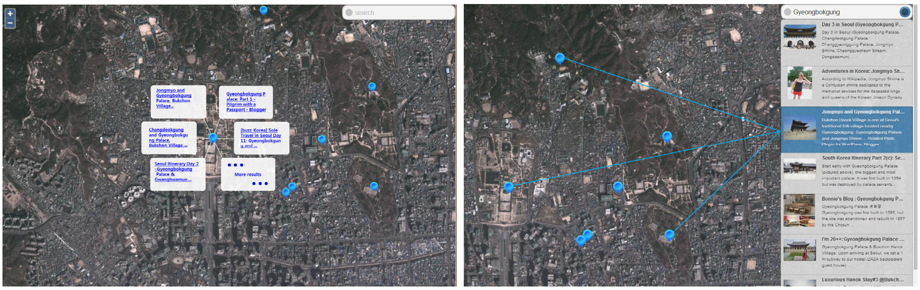
where $\boldsymbol{s}$ is the initial evidence vector, an $\mathrm{n} \times 1$ where $s_i = \mathrm{importance(m)}$ for i corresponding to m, $\boldsymbol{r^t}$ is the evidence vector at time t, an $\mathrm{n} \times 1$ vector where r_i is the evidence score for the node i and $\boldsymbol{T}$ is the evidence propagation matrix, an $\mathrm{n} \times \mathrm{n}$ matrix where T_{ij} is the evidence propagation ratio from node i to node j. We iterate $r^{t+1} = (1-\lambda) \times T \times r^t + \lambda \times s$ until the convergence. Finally we choose the highest score entity by $\mathrm{argmax}_e\, CP(m,e) \times r_d(e)$.

4.4 Multiple Geographic Focus Determination

After the POI resolution, we determine a geographic focus for the document. In previous papers[1,2,7], the only one geographic focus is emphasized in the document. Depending on the documents, this assumption cannot be true. For example, in a weblog, the writer can describe the trip the several area, thereby resulting in several different focuses. We find the multiple geographic focuses, and represent it by POI LOD, a tree structure of multiple geographic focuses. This is useful in that a geographic focus can be used for the geographic search term. For example, the search term Boston can result in documents, which have Boston LOD node.

5 Implementation

To show the feasibility of the proposed system, we have implemented a prototype system that provides a map-based search service for non-geotagged web documents. The user can input a set of words into the text input form and as a result, a set of POIs is shown in both map and list area. The given query is processed as a spatial search for web documents within the bounding box of map area. The system provides two search modes according to the search criteria, POI mode and document mode. The POI mode presents the search result by the media of POI as shown in Fig. 5(a). Therefore the user can browse the map showing the search result. The web documents are clustered around the selected POIs. If a POI marker is selected, the items signifying the retrieved web documents are shown around it. When the map is magnified, the POI is separated to a set of subordinate POIs. The system takes the user to the actual document by clicking the result items. The document mode intends to provide the optimal method for fine-grained browsing the web documents mentioning multiple POIs as shown in Fig. 5(b). In the right pane, the document list consisting of a teaser image and the summary of document is displayed in the right pane and each item is connected to the marker of specific POI. Because the multi-geographical focuses are supported, a web document can be associated to multiple POI connections.



(a) (b)

Fig. 5. Two modes of map-based web search interface. (a) shows POI mediated browsing mode of web search results. (b) shows fine-grained search for multiple POIs mentioned in a single web document.

The crawler is based on Apache Nutch 2, as its typical configuration that is combined with Hadoop, HBase, and Gora. We have developed the POI indexer in Java, as a Nutch plugin. DPOI is a separated component and provides the POI entity database functions through RESTful web service. The DPOI service has been developed with Node.js and mongoDB. The map-based search user interface is executed as a web application and mainly uses OpenLayers for map presentation.

6 Conclusion

In this paper, we presented the design and implementation of a map search interface for non geo-tagged web documents. Our system detects geographic locations based on POIs including restaurants, stores, landmarks, etc. We also determine multi-geographical focuses in the document. As a future work, we need to improve the precision and recall of toponym recognition, especially the precision. Unlike the cities, states, countries names, the POI names vary in their types. Some person's names are recognized as POIs such as McDonald's, Wendy's. In order to improve the precision, we need fine-grained NER, trained by the same POI data set. Another future work is an empirical validations described in this paper.

Acknowledgments. This research is supported in part by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2013 and by the Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 2E24790).

References

1. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 273–280. ACM, Sheffield (2004)
2. Lieberman, M.D., Samet, H., Sankaranarayanan, J., Sperling, J.: STEWARD: architecture of a spatio-textual search engine. In: Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems, pp. 1–8. ACM, Seattle (2007)
3. Vaid, S., Jones, C.B., Joho, H., Sanderson, M.: Spatio-textual indexing for geographical search on the web. In: Medeiros, C.B., Egenhofer, M., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 218–235. Springer, Heidelberg (2005)
4. Teitler, B.E., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., Sperling, J.: NewsStand: a new view on news. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 1–10. ACM, Irvine (2008)
5. Wood, J., Dykes, J., Slingsby, A., Clarke, K.: Interactive Visual Exploration of a Large Spatio-temporal Dataset: Reflections on a Geovisualization Mashup. IEEE Transactions on Visualization and Computer Graphics 13, 1176–1183 (2007)
6. Slingsby, A., Dykes, J., Wood, J., Clarke, K.: Interactive Tag Maps and Tag Clouds for the Multiscale Exploration of Large Spatio-temporal Datasets. In: 11th International Conference on Information Visualization, IV 2007, pp. 497–504 (2007)
7. Lieberman, M.D., Samet, H., Sankaranarayanan, J.: Geotagging with local lexicons to build indexes for textually-specified spatial data. In: 2010 IEEE 26th International Conference on Data Engineering (ICDE), pp. 201–212 (2010)
8. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graph-based method. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 765–774. ACM, Beijing (2011)

9. Adelfio, M.D., Samet, H.: Structured toponym resolution using combined hierarchical place categories. In: Proceedings of the 7th Workshop on Geographic Information Retrieval, pp. 49–56. ACM, Orlando (2013)
10. Tong, H., Faloutsos, C., Pan, J.-Y.: Fast random walk with restart and its applications (2006)
11. Göbel, F., Jagers, A.: Random walks on graphs. Stochastic processes and their applications 2, 311–336 (1974)
12. M Digital Marketing Report, Q3 2013 (2013)
13. 6th Annual 15miles/Neustar Localeze Local Search Usage Study Conducted by comScore (2013)
14. Kim, K.M., Oh, H.S., Lee, K.W., Kim, E.Y., Myaeng, S.H.: A Disambiguation Method for Point of Interest Detection. Journal of KIISE (in Korean)
15. Qin, T., Xiao, R., Fang, L., Xie, X., Zhang, L.: An efficient location extraction algorithm by leveraging web contextual information. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 53–60. ACM, San Jose (2010)
16. Ahern, S., Naaman, M., Nair, R., Yang, J.H.-I.: World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 1–10. ACM, Vancouver (2007)
17. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of the 15th International Conference on Multimedia, pp. 631–640. ACM, Augsburg (2007)
18. Hiramatsu, K., Kobayashi, K., Benjamin, B., Ishida, T.: Akahani, J.-i.: Map-based user interface for Digital City Kyoto. The Internet Global Summit, INET2000 (2000)