I J C T A, 9(3), 2016, pp. 1507-1516

EGGS: E-News Geo-parsing and Geocoding for Spatial Information Visualization

Naveen Kumar[1] **and S. Siva Sathya**[1]

Abstract: With the deluge of information content in the web generated through various online news portals and social networks, there is an urgent need to analyze these data in the context of geographical proximity and relatedness for various administrative and other miscellaneous purposes. Currently spatial interpretation of these information is not a direct and easy process. Often spatial or geographical references are hidden in informal, ungrammatical, and multilingual data making it cumbersome to identify and analyze the content. In this study we propose a system EGGS to crawl and extract geographical references from E-News contents. Then geocoding of extracted geographical references and visualization of news on the map to analyze the content spatially has been done. The extraction of geographical references from text consists of two phases: Location Entity extraction and Disambiguation. The main challenge is disambiguation of place names. A filtration techniques is used to resolve the issue of ambiguity. The Spatial information extracted from the text contents may be used for integrated study of text mining and spatial mining of news text.

Index Terms: Name Entity, geoparsing, geotagging, text mining, tokenize, geocoding, map visualization.

1. INTRODUCTION

The growing volume of e-news data that contains many geographical and spatial references makes it highly difficult for historians and researchers to make some useful spatial inferences from these data. People and organizations have an increasing demand for extraction of relevant information from massive amounts of news data arriving in real-time as news streams. The news content is filled with information pertaining to geographical or spatial nature, particularly location information such as addresses, postal codes and telephone numbers. This paper presents a system called EGGS for E-News Geo-parsing and geocoding and obtaining spatial knowledge inference from the huge deluge of data emanating from online news portals gathered by web-crawling programs. For each news containing geographical information, geocoding techniques are applied to identify the actual location in terms of latitude and longitude which could be plotted in a map for easy visualization. Next, it is augmented with the location information with keyword descriptors extracted from the news web page contents. Spatial data mining techniques may later be used on this augmented information to understand the spatial relatedness of the data. This allows people to confirm that the places in the news text are correctly identified and located at the time of writing.

Geoparsing is done in two steps: extraction of location named entity from text document and disambiguation is applied on extracted location names to decrease the ambiguity in the extracted named entity. Extraction of location named entity is done by processing text document by NLP Location Named Entity finding process from text document. The extraction of named entity may result in ambiguity in the location information. Disambiguation is used to associate the correct location name to the extracted named entity. These two steps together produces the location name of the associated text document that is further used in geospatial application.

In [16], CoNLL for named entity recognition is described which states that ambiguity in spatial references is bi-directional, as the same name can be utilized for more than one spatial locations (referent uncertainty),

[1] Department of Computer Science, Pondicherry University-605014, *Email: nav.bharti@gmail.com, ssivasathya@gmail.com*

and the vice-versa (reference vagueness) is also possible. Also the same name can be utilized for spatial locations as well as for persons which means, a different class of entity (referent class ambiguity). Additional directions include the incorporation of other sources of news and information. For example, incorporation of social media like Twitter tweets, Facebook posts etc. into this application resulting in the creation of an improved system.

1.1. Textual information in web pages

In the event that a geographical reference is available in a page, the extent of the page is identified with the referenced location. If a location name is present in a web page, then our context is limited to the neighborhood of that location. The primary assumption is that the more frequently a term appears in a page, the more probable it is important for that record. Having the same spatial entity occurring a few times relates to a bigger trust in the page concerning that extension. All content in a page is not of the same significance. HTML characterizes an arrangement of parts to which pieces of content can be assigned. Instinctively, terms showing up in various roles have an alternate significance (i.e. geographic references made in the title ought to be viewed as more essential, and therefor weighted in like manner). The geographical scope of a news text is related to the location of the web server where it resides [1]. Geographical references extend their influence to more than the document they occur in. Hypertext links and geographical scopes are correlated. Geographical references in hyperlink anchor text are good indicators of the scope for the target page. Some web pages are more authoritative. Hence, linkage information should be explored in an aggregate sense.

1.2. Usage of geographical references

Every news page has got a geographical scope which defines the region of interest or where the news relate to. If a news references a given spatial location then, to some degree, it ought to additionally concern its spatial sub-location, neighboring location, and related location (Autocorrelation). Demographics information can be utilized to disambiguate spatial scopes. As a matter of fact, names at higher levels will probably be referenced in site pages and are additionally more conspicuous by clients. Spatial references distinguished in site pages are more dependable than in spatial data gathered from other references. Different contexts that are in assertion give expanded certainty.

The proposed work namely EGGS is limited to E-News available in the web. The EGGS System takes the URL of news website to extract the news content with associated metadata. The extracted news is geoparsed and geocoded. These geocoded news is further used to visualize the various attribute of news. News level geoparsing begins with a user giving a keyword of his interest associated with any URL of news web portal to the system. The system starts extraction and identification of location named entity from each of the news document and prompts the user by highlighting the location information where the news happened. In case of ambiguity of location, named entity system gives provision to the user to identify and tag associated location information manually to the respective text document of news. This is a kind of geoparsing, following which, geocoding will be done to analyze the map for pattern & knowledge extraction in the later phases.

The rest of the paper is organized as follows. Related literature is reviewed in Section 2. The proposed system is described in Section 3. Section 4 describes the discussion and analysis. Section 5 concludes the paper.

2. LITERATURE REVIEW

To elaborate on the state of the art in the field of geoparsing and geolocation tagging, not much of work exists in the literature especially in the light of social network analysis [12], yet interpersonal organizations

are rapid and the information is regularly difficult to get. There are three major families [13] of strategies at present being used for the recognition of spatial spot element in content. However this study is restricted to Gazetteer Lookup Based technique. Some related work on geoparsing has been audited on two viewpoints: Location Named Entity Recognition and Location Entity Disambiguation of place names. They are reviewed briefly in this section

2.1. Location Name Entity Extraction:

In [17] and [19], a system with two modules was used for location name extraction. The Location Named Entity interface has control module and monitoring module. Text document is taken as input to the first module which discovers the associated semantics and classifies according to the identified string in the text document. This module outputs the ordered and classified strings and neglects the rest of the text which are not recognized. This module classifies the recognized strings into three classes i.e. the word with one attribute location name, location name with composite names and the words that are not location names. The input to the second module has only groups of adjacent words that are classified as location named entity. The binding module indexes the recognized named entities which are the output of the first module.

The text is then traversed either word by word or character by character, and searched for occurrences of a predefined set of toponyms. These toponyms are stored in a gazetteer, a database of place names and associated metadata [14]. Note that exceptional treatment of multi-word toponyms (e.g. “West Bengal”) might be essential, where a gullible lookup-based methodology can without much of a stretch lead to inefficiencies. If the set of toponyms is not organized in a flat list but as a hierarchy, the term ontology-based geoparsing is used.

2.2. Location Entity Disambiguation:

Extraction of location named entity and processing them is a difficult task in almost all languages because, the techniques used to present location information in the text document varies. Extraction of location named entity in the text of a document could be a challenging part because identifying words as location named entity is difficult. For this geoparsing is used that not only identifies the associated word but also classifies that that word as a location named entity. Geoparsing must understand the context of the identified place name entity and differentiate it in case of ambiguity. Ambiguity is another challenge that occur next to identification of named entity. Geoparsing may possibly identify text names that are not actually location information, and refer to the location incorrectly. These difficulties are overcome through integrating automated geoparsing techniques and location names checking system that also facilitates the user to correct location names.

The work presented in [1] is a methodology for automatically recognizing the geographic scope of the content in website pages. The procedure works in two stages. In the first stage, identification of geographical scope and disambiguation of news content is done so that the scope of the location assignment to news content is limited to a higher level of location. In the second stage, the location scope is further refined by using a page-rank algorithm which means that more the number of web pages linked to a particular geographical location, the more likely the content belongs to that location. The algorithm utilizes an ontology for assigning a geographical scope that has location as a naming entity and some relationship exist among the location. These two stages together make use of heuristics to assign the geographical scope to the news with geographical context.

In paper [4] a strategy utilizing LIW (Location Indicative Word) that has the potential for enhancing content based spatial location through element determination is used. The LIWs recognized by this technique, and their relationship with specific location is helpful for word specialists in portraying their ideas in a geographical context. The author has then considered the expectation certainty, and demonstrated that it is

conceivable to strike an exchange in the middle of scope and precision; given the massive Twitter information accessible, a framework which gives more exact forecasts. The work presented in this paper just considered tweets with highest quality level geo-tags, yet in a connected setting we imagine these models being is connected to non-geotagged tweets to deduce their spatial location. In any case, it won't not be the situation that geo-tagged tweets (normally sent from a GPS-empowered gadget, for example, an advanced smart phone) have the like properties as those which are not geo-tagged (and are sent from an assortment of gadgets, including desktop PCs). A similar work is done for classifying text according to geographic location [3], based on evidence obtained on the connections between Wikipedia entries and their titles. That showed how geographic evidence is used to build term lists that are used as classification features. The strategy for extracting geographic evidence from Wikipedia can be improved in several ways. One of them would be to expand the list of relevant terms by looking up alternative terms for the same entity. Another idea to be explored is to verify the terms used to create the hypertext links between Wikipedia entries, since they can be different from the entry's title. Furthermore, the semantic network of terms represented by Wikipedia's graph can be explored more deeply.

An OntoGazetteer [5], records semantic connections among places. This presents a next-generation gazetteer, a toponymic dictionary which expands from the traditional cataloguing of place names and includes geographic elements such as spatial relationships, concepts and terms related to places. The work used ontology concepts to define a flexible way to establish and maintain semantically richer relationships between places, and adds resources for keeping alternative names and lists of place related terms. Also techniques in which the semantically enhanced gazetteer can be used in typical geographic information retrieval tasks are presented. An integrated geolocation prediction framework presented in [6] is used to investigate the factors that impact the prediction accuracy. It first evaluates a range of feature selection methods to obtain "location indicative words", then evaluates the impact of non-geotagged tweets, language, and user-declared metadata on geolocation prediction. Also, it evaluates the impact of temporal variance on model generalization, and discusses how users differ in terms of their geolocatability. The ontological gazetteer [7] which not only identifies the names of places, but also record concepts and terms related to a place, as in an ontology in which concepts are the main places and features. This work uses ontology concepts to define a flexible way to establish and maintain semantically richer relationships between places, and adds resources for keeping alternative names and lists of place-related terms. Another ontology-driven approach is described in [8] which facilitates the process of recognizing, extracting, and geocoding partial or complete references to places embedded in text. The approach combines an extraction ontology with urban gazetteers and geocoding techniques. This ontology is used to guide the discovery of geospatial evidence from the contents of Web pages. It shows that addresses and positioning expressions, along with fragments such as postal codes or telephone area codes, provide satisfactory support for local search applications, since they are able to determine approximations to the physical location of services and activities named within Web pages.

A major problem in constructing location-based web searches is that most web-pages do not contain any explicit geocoding such as geotags. Alternative solution can be based on ad-hoc georeferencing [9], which relies on street addresses, but the problem is how to extract and validate the address strings from free-form text. To achieve this a rule-based pattern matching solution [9] is proposed that detects address-based locations using a gazetteer and street-name prefix trees created from the gazetteer.

The work in paper [11] visualizes and analyzes information ideas and events posted on public web pages through customized web-search engines and keywords. The work is the integration of GIScience and web-search engine [18] that analyze by tracking public web pages and generated spatial relationships among the pages. For this a web automatic reasoning and spatial mapping is used to search and analyze web content with provided keywords, concepts, or news etc. that generates spatial context and spatial relationships among events. This research shows an important spatial context and relationship of the searched keywords through search engine.

3. EGGS: E-NEWS GEO-PARSING AND GEOCODING FOR SPATIAL INFORMATION VISUALIZATION

This section presents the proposed system namely EGGS for geoparsing of text content from E-News. The system comprises of the following four main components: News Extractor, Geoparser, Geocoder and Visualizer as given in Figure-1.

News Extractor: The system is initialized with the URL of E-News website, Keywords of interest along with timeframe. The system learns the pattern of web page tags for extracting the content of web pages. The extractor maintains a separate file to remember the patterns in which the web pages have to be extracted and what kind of pages have to be extracted with the defined constrains. Once the pattern is identified then the component uses a web crawler to extract the news contents. Whenever a new web site is added to be crawled, first the user has to give some example URLs so that it creates the structure of news content.

Preprocessing Component: Once the web pages are extracted it gives a structure to the extracted content of the text. This component uses a Tokenizer and Concept Finder for Named Entity in the news texts and then filtration is applied to further process the tokenized texts to identify and disambiguate the place names. There are a few exclusive and open source frameworks to perceive name substance and names, and also geographic expressions. Apache's OpenNLP4 is a Java API for characteristic dialect preparing parts, including a module for named substance acknowledgment.

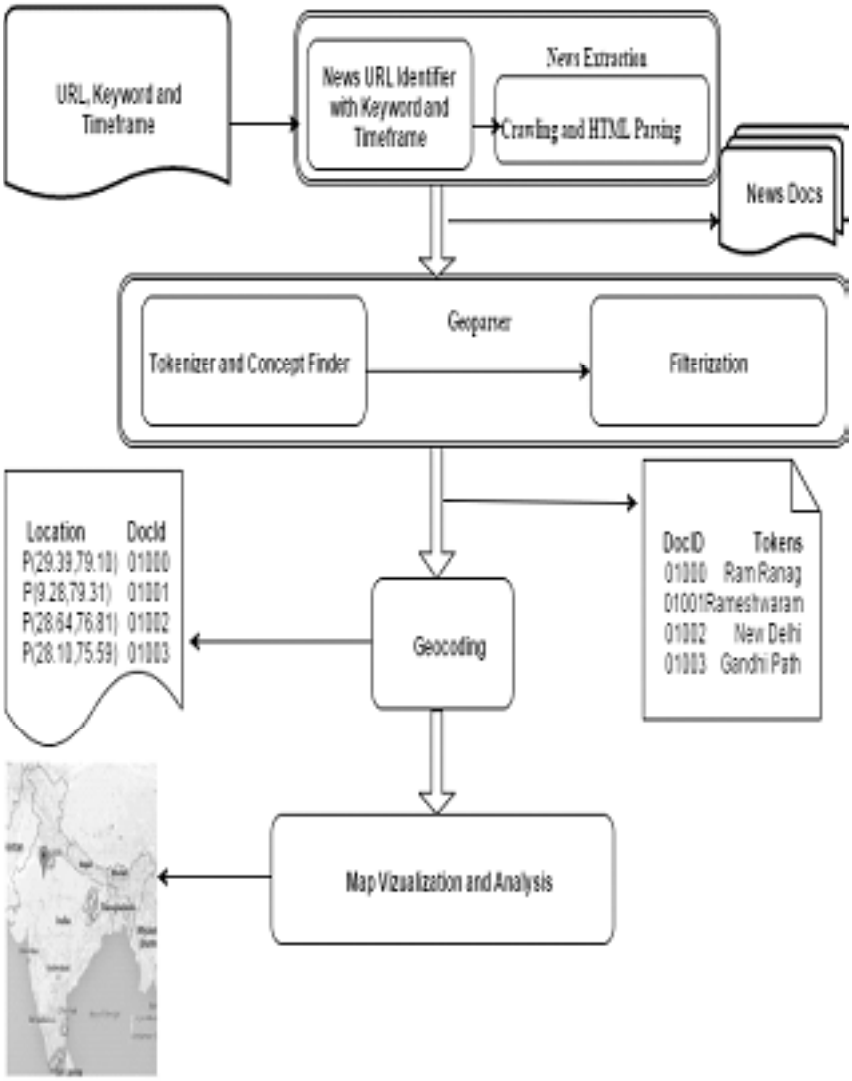

Figure 1: E-News Geoparsing And Visualization

Geocoding: The recognized geographical naming entity is assigned a weight and according to its use in the news texts, a well formed address is generated for geocoding. The resulting geocoded data is stored and associated with each news for visualization purposes. The type of geocoding information recognized is as follows: Latitude and Longitude, Country Names and City Names, Postal Code, etc.

Map Visualization: This component is for projection purpose. After Geocoding is done and assigned to news the geocoded data is used for map visualization. The map visualization gives better understanding of news. The spatial information associated to each news can be used to generate various maps. The visualization can be done in four different types of map visualizations: heat maps, cluster maps, marker maps and bubble maps.

The algorithm for EGGS is given below:

EGGS Algorithm

```
INPUT: seed URL, search Keyword, search Timeframe
    //seed URL is any news web Url.
    //Keyword is any text that will be searched in crawler.
    //Timeframe is start and end time defined as duration of news.
OUTPUT: Geocoded location of each news with keyword and Timeframe
    //The output will be geocoded location (Latitude, Longitude)
    //And visualization map with place marker.
BEGIN
    INTIALIZE: URLs; Key; Timeframe(startTime, endTime)
    //The parameters are initialized with the given input.
    UrlStrct = trainPagePttern (with example url)
    //Url Structure is indentified using trainPagePattern with given news urls.
    URLSet = CrawlURL(URL, Keyword, Timeframe)
    //A set of Urls are crawled within defined Timeframe and Keyword.
    Initialize NewsItem, NewsLoc, AddLocSet to null
    FOR each url in URLSet DO
        //For each Url in URLSet Parse News text,
        //Geoparse place name, and Geocode.
        NewsItem = Parse(url)
        NewsLoc = GeoParse(NewsItem)
        AddLocSet = Geocode(NewsLoc)
END FOR
mapBox = mapFrame(AddLoc.minPoint, AddLoc.MaxPoint)
//create mapframe with upper-left and Lower-right Point
mapBox.Plot(AddLocSet, 'Marker')
```

```
//A map visualization is plotted of place marker with news title.
END EGGS
PROCEDURE GeoParse(NewsItem)
    //INITIALIZE place name search Domain geoDomain place name in news heading;
    geoDomain = GeoDomainIdentifier(NewsItem[Head])
    tokens = Tokenization(NewsItem)
    //tokenize the newsitems into tokens or blocks
    //PlaceNameEntityModel is classical model for place name identification
    placeNameModel = setPatternRule(PlaceNameEntityMolde)
    //Set the rule for place name identifier Model
    placeDisambiguattion = setPatternDisambiguation(PlaceNameEntityModel, geoDomain)
    //Set Disambiguation rule for placeNameModel
    placeNames = placeNameIdentifier(tokens, PlaceNameModel)
    placeNames.Distambiguate(placeDisambiguattion)
    //Apply disambiguation on identified placeNames and
    //Identify place name entity and assigns weight
    RETURN(placeNames)
END PROCEDURE
```

4. EXPERIMENTAL SETUP & RESULTS

The system is implemented in Java platform and used OpenStreetMap map services for map visualization. The URL of news website (i.e., Times of India and Dainik Jagaran archives) have been tested for Indian

Figure 2: Maps of PM Visit and News Source

region. The News Extractor component is found to detect all the Hypertext Link and extract 95% URL of the given parameters. The 5% error which is in the extraction process is evaluated and found that those URL which are not extracted by the extractor is because of unavailability of web-pages or server response time. Once the spatial information is extracted from news content, the geocoding is done using map geocoding services, and found that most of the local geographical names could not be recognized, specially village, street, House Number, Land Mark etc. A hierarchy (i.e., Country, State, District, Sub-District, Village, Street, Land Mark etc.) is followed to correctly recognize place naming words in geocoding where the top level of hierarchy is Country and the bottom level of hierarchy is Land Mark. We have used an upper level geographical name where local names are not recognized with confidence as shown in the Figure-2. Also a

Figure 3: Place Name Word Cloud

Table 1
News Extracted Each Day

News Identified Each Day		*News Identified Each Day*	
Day (Dec 2015)	*No. of Location*	*Day (Dec 2015)*	*No. of Location*
1	544	13	549
2	523	14	556
3	533	15	591
4	522	16	564
5	505	17	629
6	536	18	580
7	540	19	599
8	592	20	582
9	563	21	552
10	572	22	548
11	582	23	540
12	565	24	170

Table 2
News Geocoding to Each City

News Identified at Each Place		*News Identified at Each Place*	
Place Name	*No. of News*	*Place Name*	*No. of News*
NEW DELHI	1132	CHANDIGARH	183
MUMBAI	594	NAGPUR	146
CHENNAI	434	NASHIK	130
Nagpur	428	KOCHI	129
LUCKNOW	407	Mumbai	127
HYDERABAD	275	PATNA	119
AHMEDABAD	263	WASHINGTON	119
BENGALURU	252	VADODARA	117
KOLKATA	252	GURGAON	116
New Delhi	229	AURANGABAD	107
NAVI MUMBAI	222	BHUBANESWAR	107
Panaji	221	JAIPUR	107
PUNE	210		

word-cloud is generated which show the most found places (i.e. Mumbai, Chandigarh, etc.) from news content shown in the Figure-3. The resulting Figure-2 of place mark map shows the geographical locations of the given keyword input “PM Modi visited” in the timeframe of December 2015 to January 2016.

The table-1 lists the number of news extracted each day with location information and table-2 shows the number of news pertaining to each city in India with respect to the given keyword and URL. This shows the volume of information that is being processed by EGGS in order to geocode and geoparse the e-news information for extracting the geographical context of the required keywords.

5. CONCLUSION

Currently, spatial interpretation and inference of news data produced massively in the web is not a direct and easy process. Visualizing these information in the context of geographical proximity or geocoding the information in a map adds more meaning to the extracted information. Hence, this paper has proposed a system to geoparse the E-News to extract the news website pages and gives structure to the free textual format of news contents. The extracted news text are used to extract place name entity and disambiguate the place names to generate a well-structured address format for the accuracy of geocoded value. This could help in many domain of interesting applications like journalists and social science researchers exploring newspapers, for historians working with historical documents, for analysis of biographies etc. It could help in quick understanding of events geographically by news texts. The visualization tool gives potential for assisting in digital humanities: at a glance the researcher gains an idea of the spatial interaction described in the text file. In future, plans exist to do temporal study and spatial story telling by using machine learning techniques and text association mining.

References

[1] Silva, Mário J., et al. “Adding geographic scopes to web resources.”*Computers, Environment and Urban Systems* 30.4 (2006): 378-399.

[2] Amitay, Einat, et al. “Web-a-where: geotagging web content.” Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.

[3] Odon de Alencar, Rafael, Clodoveu Augusto Davis Jr, and Marcos André Gonçalves. “Geographical classification of documents using evidence from Wikipedia.” *proceedings of the 6th Workshop on geographic information retrieval*. ACM, 2010.

[4] Bo, Han, and Paul COOK1 Timothy BALDWIN. "Geolocation prediction in social media data by finding location indicative words." *Proceedings of COLING 2012: Technical Papers* (2012): 1045-1062.

[5] Machado, Ivre Marjorie R., et al. "An ontological gazetteer and its application for place name disambiguation in text." *Journal of the Brazilian Computer Society* 17.4 (2011): 267-279.

[6] Han, Bo, Paul Cook, and Timothy Baldwin. "Text-based twitter user geolocation prediction." *Journal of Artificial Intelligence Research* (2014): 451-500.

[7] Machado, Ivre Marjorie, et al. "An Ontological Gazetter for Geographic Information Retrieval." *GeoInfo*. 2010.

[8] Borges, Karla AV, et al. "Ontology-driven discovery of geospatial evidence in web pages." *GeoInformatica* 15.4 (2011): 609-631.

[9] Tabarcea, Andrei, Ville Hautamäki, and Pasi Fränti. "Ad-hoc georeferencing of web-pages using street-name prefix trees." *Web Information Systems and Technologies*. Springer Berlin Heidelberg, 2011. 259-271.

[10] Cardoso, Nuno. "Rembrandt-a named-entity recognition framework." *LREC*. 2012.

[11] Samet, Hanan, et al. "Reading News with Maps: The Power of Searching with Spatial Synonyms."

[12] Backstrom, Lars, Eric Sun, and Cameron Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity."*Proceedings of the 19th international conference on World Wide Web*. ACM, 2010.

[13] Leidner, Jochen L., and Michael D. Lieberman. "Detecting geographical references in the form of place names and associated spatial natural language." *SIGSPATIAL Special* 3.2 (2011): 5-11.

[14] Hill, Linda L. Georeferencing: The geographic associations of information. Mit Press, 2009.

[15] Hopcroft, John E. Introduction to automata theory, languages, and computation. Pearson Education India, 1979.

[16] Tjong Kim Sang, Erik F., and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition."*Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003.

[17] Qin, Teng, et al. "An efficient location extraction algorithm by leveraging web contextual information." *proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2010.

[18] Tsou, Ming-Hsiang, et al. "Mapping ideas from cyberspace to realspace: visualizing the spatial context of keywords from web page search results."*International Journal of Digital Earth* 7.4 (2014): 316-335.

[19] Jevtic, Dragan, Zeljka Car, and Marin Vukovic. "Location name extraction for user created digital content services." *Knowledge-Based Intelligent Information and Engineering Systems*. Springer Berlin/Heidelberg, 2007.