Opinion Mapping Travelblogs

Efthymios Drymonas, Alexandros Efentakis, and Dieter Pfoser

Institute for the Management of Information Systems
Research Center Athena
G. Mpakou 17, 11524 Athens, Greece
{edrimon|efentakis|pfoser}@imis.athena-innovation.gr

Abstract. User-contributed content represents a valuable information source provided one can make sense of the large amounts of unstructured data. This work focusses on geospatial content and specifically on travelblogs. Users writing stories about their trips and related experiences effectively provide geospatial information albeit in narrative form. Identifying this geospatial aspect of the texts by means of applying information extraction techniques and geocoding, we relate portions of texts to locations, e.g., a paragraph is associated with a spatial bounding box. To further summarize the information, we assess the opinion ("mood") of the author in the text. Aggregating this mood information for places, we essentially create a geospatial opinion map based on the user-contributed information contained in the articles of travelblogs. We assessed the proposed approach with a corpus of more than 150k texts from various sites.

1 Introduction

Crowdsourcing moods and in our specific case opinions from user-contributed data, has recently become an interesting field with the advent of micro-blogging services such as, e.g., Twitter. Here, blog entries reflect a myriad of different user opinions that when integrated can give us valuable information about, e.g., the stock market [4]. In this work, our focus is on (i) extracting the user opinion about places from travel blog entries, (ii) aggregating such opinion data, and, finally, (iii) visualizing it.

The specific contributions in this work are as follows. In an initial stage several travelblog Web sites have been crawled and over 150k texts have been collected. Figure 5 shows such an example travelblog entry. The collected texts are then geoparsed and geocoded to link placename identifiers (toponyms) to location information. With paragraphs as the finite granularity for opinion information, texts are then assessed with the OpinionFinder tool and are assigned a score for each paragraph ranging from very negative to very positive. Scores are linked to the bounding box of the paragraph and are aggregated using a global grid, i.e., the score of a specific paragraph is associated with all intersecting grid cells. Aggregation of opinions is then performed simply by computing the average of all scores for each cell. Finally, the score can be visualized by assigning colors to each cell.

While to the best of our knowledge there exists no work aiming at extracting opinions about places from travel blogs, we can cite the following related work. The concept of information visualization using maps is gaining significant interest in various research fields. As examples, we can cite the following works [26] [23] [29] [12]. For the purpose of recognizing toponyms, the various approaches use ideas and work from the field of Natural Language Processing (NLP), Part-Of-Speech (POS) tagging and a part of Information Extraction related tasks, namely Named Entity Recognition (NER) [13]. These approaches, can be roughly classified as rule-based [6] [7] [8] [21] [30] and machine learning - statistical [14] [16] [26] [22]. Once toponyms have been recognized, a toponym resolution procedure resolves geo-ambiguity. There are many methods using a prominence measure such as population combined with other approaches [21] [25]. With respect to geocoding, we can exemplary cite [17], one of the first works on geocoding and describing a navigational tool for browsing web resources by geographic proximity as an alternative means for Web navigation. Web-a-Where [1] is another system for geocoding Web pages. It assigns to each page a geographic focus that the page discusses as a whole. The tagging process targets large collections of Web pages to facilitate a variety of location-based applications and data analyses. The work presented in [15] is identifying and disambiguating references to geographic locations. Another method that uses information extraction techniques to geocode news is described in [26]. Other toponym resolution strategies involve the use of geospatial measures such as minimizing total geographic coverage [14], or minimizing pairwise toponym distance [16]. An approach for the extraction of routes from narratives is given in [9]. The proposed IE approach has been adapted to fit the requirements of this work. While statistical NER methods can be useful for analysis of static corpora, in the case of continuously user contributed travel narratives they are not well-suited, due to their dynamic and ever-changing nature [25]. For this purpose, we rely on a powerful rule-based solution based on a modular pipeline of distinct, independent and well-defined components based on NLP and IE methods, as we will see in the next section. Regarding related work on opinion classification and sentiment analysis [20], we can find methods basically relying on streaming data [18] [10] [11] [19]. Recently [2] discusses the challenges that Twitter streaming data poses. The work focusses on sentiment analysis and proposes the sliding-window Kappa statistic as an evaluation metric for data streams.

The remainder of this work is organized as follows. Section 2 describes the information extraction techniques employed in our approach dealing specifically with the aspects of geoparsing and geocoding travel blog entries. Section 3 outlines a method for computing user sentiment scores from travel blog entries. Section 4 outlines how such scores can be aggregated and visualized based on geospatial locations. In addition some specific examples are shown to give an initial validation of the proposed approach. Finally, Section 5 presents conclusions and directions for future work.

2 Information Extraction

In what follows, we describe in detail the processing pipeline, which overall uses an HTML document as input (travel blog article) and produces a structured XML file containing the various entities and their respective attributes (toponyms and coordinate information for the text).

The pipeline consist of four parts (cf. Figure 1), (i) the HTML parsing module, (ii) the linguistic pre-processing, (iii) the main IE engine system (semantic analysis) and

(iv) the geocoding-postprocessing part. In the next section, we describe the first part of our processing pipeline, i.e., the collection of HTML texts, the parsing and their conversion to plain text format, in order to prepare the documents for the forthcoming step of linguistic-preprocessing.

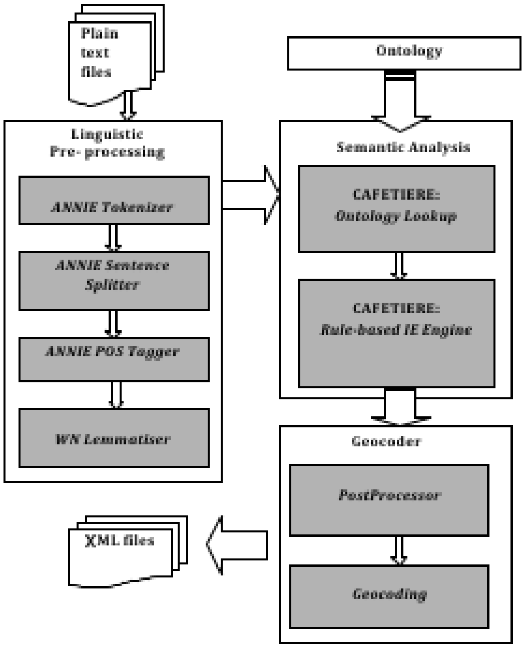


Fig. 1. IE architecture pipeline

2.1 Web Crawling

For collecting travel blog articles containing rich geospatial information, we crawled Web sites providing traveblog authoring services. Each Web site has its own HTML layout and isolating text of interest from crawled and parsed HTML pages is done by hand. Thus, there was a need for Web sites with massive amounts of such type of documents. For this purpose we crawled travelpod.com, travelblog.org, traveljournals.net and worldhum.com, resulting in more than 150,000 documents. For crawling the web sites, we used Regain crawler[1], which creates a Lucene[2] index for indexing the documents' information, while for HTML parsing and the extraction of useful plain text narratives, we used Jericho HTML parser[3].

[1] http://regain.sourceforge.net/
[2] http://lucene.apache.org/
[3] http://jericho.htmlparser.net/

2.2 Linguistic pre-Processing

To prepare the input for the core IE engine for extracting objects of interest, the parsed plain text documents must be prepared accordingly. Such preparation includes linguistic pre-processing tools that analyze natural language documents in terms of distinct base units (i.e., words), sentences, part-of-speech and morphology . We are using the ANNIE tools, contained in the GATE release[4], to perform this initial part of analysis. To this task, our processing pipeline comprises of a set of four modules: (i) the ANNIE tokenizer, (i) the (ANNIE) Sentence Splitter, (iii) the ANNIE POS Tagger and (iv) the WordNet Lemmatiser.

The intermediate processing results are passed on to each subsequent analysis tool as GATE document annotation objects. The output of this analysis part is the analyzed document and it is transformed in CAS/XML format[5], which will be passed to the subsequent semantic analysis component as input, Cafetiere IE engine [3]. Cafetiere combines the linguistic information acquired by the pre-processing stage of analysis with knowledge resources information, namely the lookup ontology and the analysis rules to semantically analyze the documents and recognize spatial information, as we will see later in this section.

The first step in the pipeline process is *tokenization*, i.e., recognizing in the input text basic text units (tokens), such as words and punctuation and orthographic analysis and the association of orthographic features, such as capitalization, use of special characters and symbols, etc. to the recognized tokens. The tools used are ANNIE Tokenizer and Orthographic Analyzer.

Sentence splitting, in our case the ANNIE sentence splitter aims at the identification of sentence boundaries in a text.

Part-of-speech (POS) tagging is then the process of assigning a part-of-speech class, such as Noun, Verb etc. to each word in the input text. The ANNIE POS Tagger implementation is a variant of Brill Transformation-based learning tagger [5], which applies a combination of lexicon information and transformation rules for the correct POS classification.

Lemmatisation is used for text normalisation purposes. With this process we retrieve the tokens base form e.g., for words: [travelling, traveler, traveled], [are, were], the corresponding lemmas are: travel, be. We exploit this information in the semantic rules section. For this purpose we implement the JWNL WordNet Java Library API[6] for accessing the WordNet relational dictionary. The output of this step is included it in GATE document annotation information.

2.3 Semantic Analysis

Semantic analysis relates the linguistic pre-processing results to ontology information, as we will see in the next subsection about ontology lookup and applies semantic analysis grammar rules, i.e., documents are analyzed semantically to discover spatial concepts and relations.

[4] http://gate.ac.uk/

[5] CAS is an XML scheme called Common Annotation Scheme allowing for a wide range of annotations, structural, lexical, semantic and conceptual.

[6] http://sourceforge.net/projects/jwordnet/

For this purpose we used Cafetiere IE engine, whose objective is to compile a set of semantic analysis grammar rules in a cascade of finite state transducers so as to recognize in text the concepts of interest. Cafetiere IE Engine combines all previously acquired linguistic and semantic information with contextual information. We modified Cafetiere and implemented it as a GATE pipeline module (GATE creole) for the purpose of performing ontology lookup and rule-based semantic analysis on information acquired from previous pipeline modules, in the form of GATE annotation sets. The input to this process are the GATE annotation objects resulted from the linguistic preprocessing stage stored transformed in Cafetiere needed format, in CAS/XML format for each individual document.

Cafetiere Ontology Lookup The use of knowledge lexico-semantic resources assists in the identification of named entities. These semantic knowledge resources may be in the form of lists (gazetteers) or more complex ontologies providing mappings of text strings to semantic categories, such as in general male/female person names, known organizations and known identifiers of named entities. In our case, the named entities we want to extract with IE methods are location based information. For example, a gazetteer for location designators might have entries such as "Sq.", "blvd.", "st." etc. that denote squares, boulevards and streets accordingly. Similarly there are various sorts of gazetteers available for given person names, titles, location names, companies, currencies, nationalities etc. Thus, the named entity (NE) recognizer can use gazetteer information so as to classify a text string as denoting an entity of a particular class. However, in order to associate specific individual entities object identifiers are required as well as class labels, enabling aliases or abbreviations to be mapped to a concrete individual. For example, for an entity such as "National Technical University of Athens" the respective abbreviation "NTUA" could be included in the knowledge resource as an alias for the respective entity. Thus, more sophisticated knowledge resources than plain gazeteers in the form of ontologies may be used to provide this type of richer semantic information and allow for the specification and representation of more information, if necessary, than identity and class inclusion.

In this way, Cafetiere Ontology lookup module accesses a previously built ontology to retrieve potential semantic class information for individual tokens or phrases. All types of conceptual information, related to domain specific entities, such as terms or words in general that denote spatial concepts or properties and relations of domain interest are pre-defined in this ontology. For example, consider the partial ontology shown in Figure 2. Class "LOCVERB" stores verbs that when matched to a text phrase are likely to indicate a spatial relationship between the corresponding referenced concepts. We label as semantic any classification of tokens according to their meaning in the field of the application, in our case, geosemantics. This could be done, on a broad coverage level, by reading information from a comprehensive resource such as WordNet lexicon about most content words. However, the practice in information extraction applications as discussed in previous paragraph, has been to make the processing application-specific by using lists of the semantic categories of only relevant words and phrases, done by hand. The ontology used in our experimentation was created by manually analyzing a large number of texts and iteratively refining the ontology with words (e.g., verbs) that

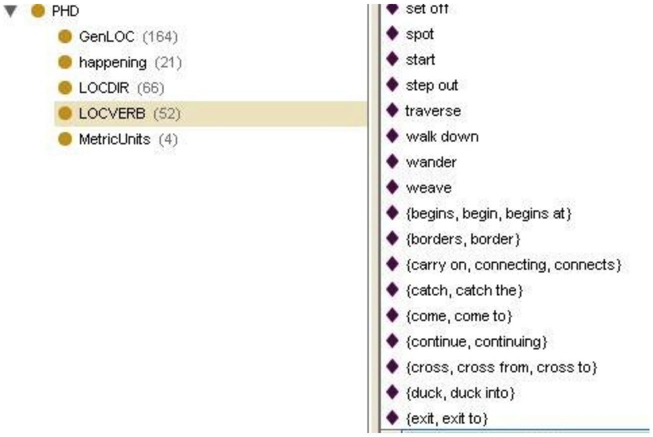


Fig. 2. Sample ontology contents (Protg ontology editor)

when matched to a text phrase are likely to indicate a spatial relationship between the corresponding referenced concepts. Summarizing, the lookup stage of analysis:

1. Supplies semantic classes (concepts) corresponding to words and phrases.
2. Supplies object identifiers for known instances, including where aliases and abbreviations name the same instance (For example “National Technical University of Athens”, “NTUA”).
3. Supplies properties of known instances, for example the country of which a city is the capital.
4. Uses verbs of interest to the application in order to identify inside the phrase potential unknown instances.

Cafietiere Information Extraction engine The approaches to Named Entity recognition with IE methods can be divided into two main categories:

- Linguistic/rule-based approaches: in these approaches the Named Entity recognition is based on linguistic/semantic rules defining the possible linguistic patterns denoting Named Entity concepts, such as for example the approaches adopted by ANNIE[7], and Cafetiere [3]. These approaches can achieve better results than most statistics or machine learning approaches, but they require extensive human effort for the development of the necessary knowledge resources (rules and lexico-semantic resources, like ontologies, described in Cafetiere ontology lookup section). For this reason the adaptation of rule-based systems to new domains is a slow and laborious process.

[7] http://gate.ac.uk/

– Machine learning/statistics-based approaches: these approaches view Named Entity recognition as a classification problem and, they have gained increased popularity due to their relatively rapid development/ domain customization and the reduced amount of human effort required.

Cafetiere is a rule-based system for IE. A set of linguistic patterns (i.e., extraction rules) is written taking into account the lookup ontology and all previously acquired information from linguistic pre-processing. The semantic analysis rules, are developed as a set of context-sensitive/context-free grammar (CSG/CFG) rules and are compiled in a cascade of finite state transducers so as to recognize the concepts of interest in plain texts.

2.4 PostProcessing

In this part, all information regarding each object of interest for each document in the collection is imprinted as a GATE annotation set object. For each document, we have collected information about all extracted entities, along with their respective paragraph, sentence and character offset in this document. During the HTML parsing process we keep the scope that each document is referred to in order to use this information for geocoding each extracted entity. For geocoding, we initially implemented YAHOO! Placemaker[8] and used in combination with Cafetiere's output, in order to deliver better results. We observed that PlaceMaker worked well for disambiguating some entities, but it identified significant fewer place entities than our IE engine. Thus, in the remaining entities extracted by Cafetiere, we applied YAHOO! Placefinder[9] to geocode this place information passing the scope information described below for delivering more accurate results.

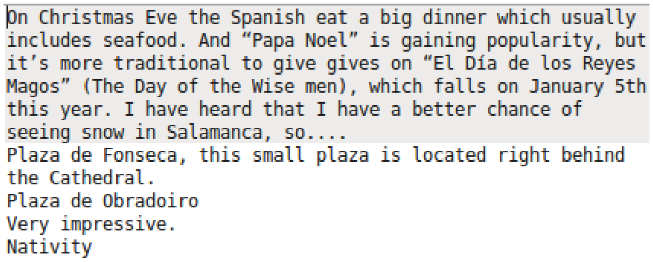
Finally, for each HTML travel blog entry (narrative), we created a collection of extracted referred geo-entities, some of them not being able to geocode. For each of these entities there is specific information (acquired from each of the previous pipeline steps) about where they were encountered in the respective document, namely, sentence, paragraph and offset character. Additionally, for each document, we calculated the mean coords and standard distance from all geocoded points extracted. All this information, along with the local parsed text file path and the respective URL of the document, are stored lastly into XML format for each corresponding plain text narrative. Samples of plain text narrative and the corresponding structured XML file are shown in Figure 3 and Figure 4 respectively. The XML tags in Figure 4 are denoting either statistical information, like the mean center and the standard distance of all geocoded locations for each document, or information related with each extracted entity, i.e., the offset characters, the sentence and paragraph ID.

3 Opinion Mapping

Having geocoded the travel blog entries, we, in the following step, want to assign sentiment information ("mood") to text. To this effect, we use OpinionFinder [28], a sys-

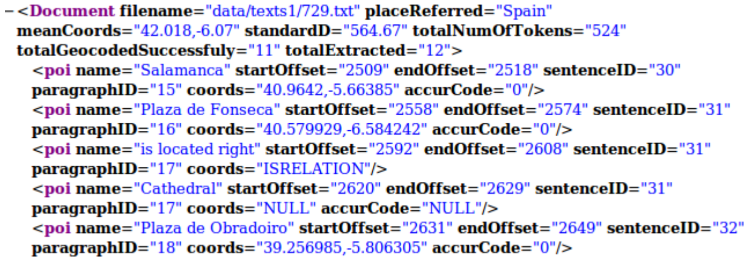
[8] http://developer.yahoo.com/geo/placemaker/

[9] http://developer.yahoo.com/geo/placefinder/



```
On Christmas Eve the Spanish eat a big dinner which usually
includes seafood. And “Papa Noel” is gaining popularity, but
it’s more traditional to give gives on “El Día de los Reyes
Magos” (The Day of the Wise men), which falls on January 5th
this year. I have heard that I have a better chance of
seeing snow in Salamanca, so....
Plaza de Fonseca, this small plaza is located right behind
the Cathedral.
Plaza de Obradoiro
Very impressive.
Nativity
```

Fig. 3. Sample plain text



```
–<Document filename="data/texts1/729.txt" placeReferred="Spain"
  meanCoords="42.018,-6.07" standardD="564.67" totalNumOfTokens="524"
  totalGeocodedSuccessfuly="11" totalExtracted="12">
    <poi name="Salamanca" startOffset="2509" endOffset="2518" sentenceID="30"
    paragraphID="15" coords="40.9642,-5.66385" accurCode="0"/>
    <poi name="Plaza de Fonseca" startOffset="2558" endOffset="2574" sentenceID="31"
    paragraphID="16" coords="40.579929,-6.584242" accurCode="0"/>
    <poi name="is located right" startOffset="2592" endOffset="2608" sentenceID="31"
    paragraphID="17" coords="ISRELATION"/>
    <poi name="Cathedral" startOffset="2620" endOffset="2629" sentenceID="31"
    paragraphID="17" coords="NULL" accurCode="NULL"/>
    <poi name="Plaza de Obradoiro" startOffset="2631" endOffset="2649" sentenceID="32"
    paragraphID="18" coords="39.256985,-5.806305" accurCode="0"/>
```

Fig. 4. Resulting XML file

tem that performs *subjectivity analysis*, automatically identifying when opinions, sentiments, or speculations are present in text. It aims to identify subjective sentences, as also marking various aspects of subjectivity in these sentences, including the source (holder) of the subjectivity and words that are included in phrases expressing positive or negative sentiments.

OpinionFinder operates as one large pipeline. Conceptually, the pipeline can be divided into two parts. The first part performs mostly general purpose document processing (e.g., tokenization and part-of-speech tagging). The second part performs the subjectivity analysis. The results of the subjectivity analysis are returned to the user in the form of SGML/XML markup of the original documents.

For the first part, OpinionFinder takes any incoming text source and removes HTML or XML meta info. Sentences are split and POS tagged using OpenNLP[10], the open source solution providing a variety of java-based NLP tools which perform sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and coreference using the OpenNLP Maxent machine learning package. Next, stemming is accomplished using Steven Abneys’ SCOL v1K stemmer program[11]. SUNDANCE (Sentence UNDerstanding And ConceptExtraction) [28], is used to provide semantic

[10] http://opennlp.sourceforge.net/

[11] http://www.vinartus.net/spa/

class tags, identify extraction patterns needed by the sentence classifiers, identifying the source of subjective content and distinguishing author statements from related or quoted statements. A final parse in batch mode establishes constituency parse trees, which are converted to dependency parse trees for Named Entity and subject detection.

At this point, for the second part, a Naive Bayes classifier identifies subjective sentences. The classifier is trained against subjective and objective sentences generated by two additional rule-based classifiers drawing from large corpora [27]. Next, a direct subjective expression and speech event classifier tags the direct subjective expressions and speech events found within the document using WordNet[12]. The final step applies actual sentiment analysis to sentences that have been identified as subjective. This is accomplished with two classifiers that were developed using the BoosTexter [24] machine learning program and trained on the MPQA Corpus[13].

4 Mapping Opinion Scores

OpinionFinder produces sentiment information assigned to paragraphs of texts. In the following, we describe how this information can be aggregated for specific locations.

4.1 Aggregating Sentiments

OpinionFinder was applied to all texts of our collection of 150k travel blog entries assigning sentiment data to each paragraph of the collection. In the analysis that follows, only paragraphs containing geospatial data were retained. For each of these paragraphs we keep the total referred positive and negative sentiment scores as computed by OpinionFinder.

Each paragraph contains zero, one or multiple geographic entities that were suitably geocoded. In order to show the spatial extent of a paragraph, we chose to spatially visualize only paragraphs in which the MBR of the contained toponyms does not exceed 0.5 degrees in either dimension (e.g., $Max\ latitude - Min\ latitude \leq 0.5$ AND $Max\ longitude - Min\ longitude \leq 0.5$). Consequently, only paragraphs of limited and focused spatial extent are visualized, thus preventing paragraphs that refer to larger geographic entities (e.g., Europe) to dominate in the results.

We used five different categories for mapping opinion scores. The categories and respective color are given in Table 1, where each category scales from negative (red) to positive (green).

The proposed approach is clarified by the following example. A sample document[14] (Figure 5) contains several paragraphs mentioning Washington D.C. and its landmarks. For each of this document's paragraphs, a MBR covering the discovered toponyms was created and each paragraph was assigned a category according to Table 1. Therefore, this document may be spatially visualized on a map as shown on Figure 6.

Although this approach is viable when there is a limited number of documents and paragraphs, we need to overcome the following problem. Multiple paragraphs from different documents and different scores may partially target the same area, e.g., we need

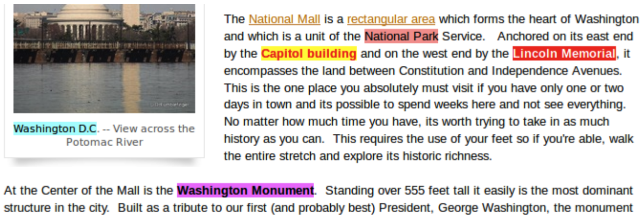
[12] http://wordnet.princeton.edu/

[13] http://nrrc.mitre.org/NRRC/publications.htm.

[14] http://www.travelpod.com/travel-blog-entries/drfumblefinger/1/1269627251/tpod.html

Result (positive - negative)	Colour
≤ -3	Red
=-1 OR =-2	Orange
0	Yellow
=1 Or =2	Olive
≥ 3	Green

Table 1. Opinion mapping to colour representation



Washington D.C. -- View across the Potomac River

The National Mall is a rectangular area which forms the heart of Washington and which is a unit of the National Park Service. Anchored on its east end by the Capitol building and on the west end by the Lincoln Memorial, it encompasses the land between Constitution and Independence Avenues. This is the one place you absolutely must visit if you have only one or two days in town and its possible to spend weeks here and not see everything. No matter how much time you have, its worth trying to take in as much history as you can. This requires the use of your feet so if you're able, walk the entire stretch and explore its historic richness.

At the Center of the Mall is the Washington Monument. Standing over 555 feet tall it easily is the most dominant structure in the city. Built as a tribute to our first (and probably best) President, George Washington, the monument

Fig. 5. Washington D.C. - sample document and toponyms

to visualize partially overlapping MBRs with different scores (colors). To do that, we split each paragraph MBR into small cells of a regular grid of 0.0045 degrees (corresponding to 500m) in each dimension. For each of those cells we sum up the sentiment score from all the containing paragraph MBRs. With this approach, instead of trying to visualize *overlapping* paragraph MBRs with different scores (colors), we visualize *distinct* small cells with each being assigned a unique score (and color). Consequently, it is easy to visualize the overall sentiment scores independent of how many paragraphs target the same area.

4.2 Opinonmap Examples

Further examples shown in the following include the geospatial opinion map of Amsterdam of Figure 7. It is interesting to observe that while most of the city is shaded green, the area around the train station and the Red Light district are shown in red, i.e., expressing rather negative sentiment.

Figure 8 gives a geospatial opinion map of Central Europe indicating the areas mentioned in the travel blogs. What can be observed is that positive sentiments are associated with areas in Switzerland and also Italy, while urban areas such as Brussels overall attract more negative sentiments.

4.3 Summary

Our initial experiments with the creation of geospatial opinion maps derived from subjective travelblog entries show that there is a clear bias for certain geographic areas shared by people. However, since in this work we only performed a simple aggregation of the scores generated by the OpinionFinder tool, it will require more in-depth analysis of the results to generate accurate statements and trends.

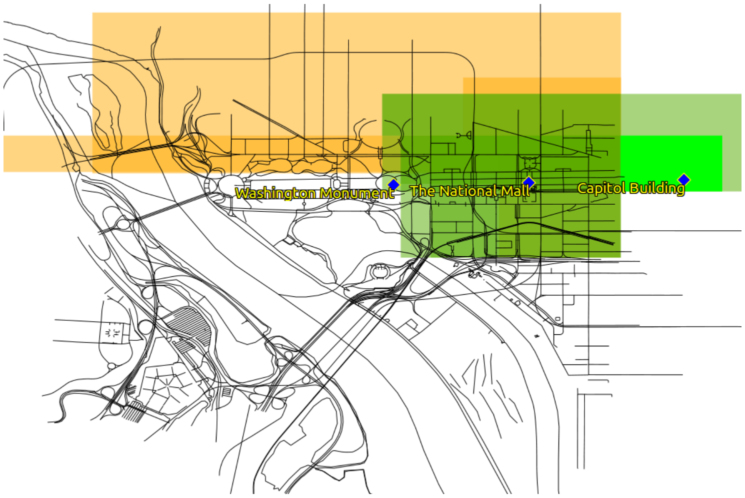


Fig. 6. Washington D.C. - geospatial opinion visualization

5 Conclusions

Aggregating opinions is important for utilizing and assessing user-generated content. This work provides a means of visualizing sentiments for specific geographic areas as derived from travel blog entries. To demonstrate the approach, several travel blog sites were crawled and a total of more than 150,000 pages/articles were processed. Using (i) geoparsing and geocoding tools the content was geo referenced and (ii) sentiment information was derived using the OpinionFinder tool. In the proposed approach, sentiment information from various articles relating to the same geographic area is aggregated and visualized accordingly by means of a geospatial heat meat. Directions for future work are as follows. The current approach for aggregating user sentiment for geographic areas is rather simple and a more in-depth analysis of the results is needed to generate accurate statements and trends. An obvious improvement will also be to examine/include microblogging content streams. Here, sentiment information will be updated live and thus represent an accurate picture of the situation of a specific geographic area over time. Finally, OpinionFinder is a general purpose tool for deriving user sentiment. More involved approaches exist and need to be examined/developed for the case of geospatial data.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme - Marie Curie Actions, Initial Training Network GEOCROWD (http://www.geocrowd.eu) under grant agreement No. FP7-PEOPLE-2010-ITN-264994.

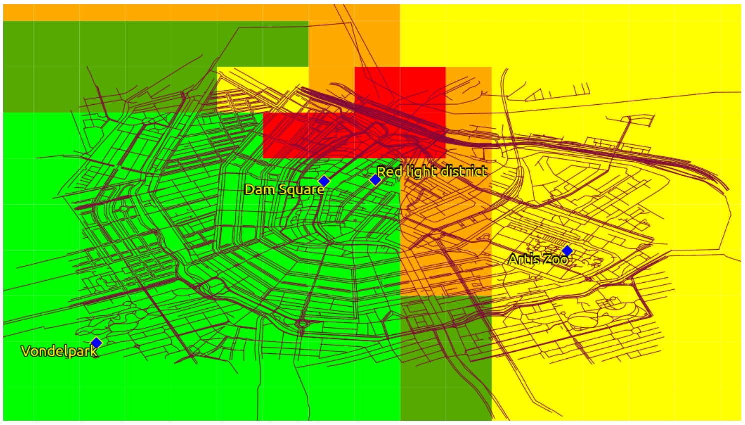


Fig. 7. Amsterdam, The Netherlands - geospatial opinion visualization

References

1. E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 273–280, New York, NY, USA, 2004. ACM.
2. A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th international conference on Discovery science*, DS'10, pages 1–15, Berlin, Heidelberg, 2010. Springer-Verlag.
3. W. J. Black, J. McNaught, A. Vasilakopoulos, K. Zervanou, B. Theodoulidis, and F. Rinaldi. Cafetiere: Conceptual annotations for facts, events, terms, individual entities and relations. Technical report, Jan 2005. Parmenides Technical Report, TR-U4.3.1.
4. J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
5. E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21:543–565, 1995.
6. D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, 22:301–313, January 2008.
7. P. Clough. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval*, GIR '05, pages 25–30, New York, NY, USA, 2005. ACM.
8. J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 545–556, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
9. E. Drymonas and D. Pfoser. Geospatial route extraction from texts. In *DMG '10: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics*, pages 29–37, New York, NY, USA, 2010. ACM.
10. A. Go, R. Bhayani, and L. Huang. Twitter Sentiment Classification using Distant Supervision. Technical report, Stanford University.
11. B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. In *Proceedings of the 27th international conference extended abstracts on*

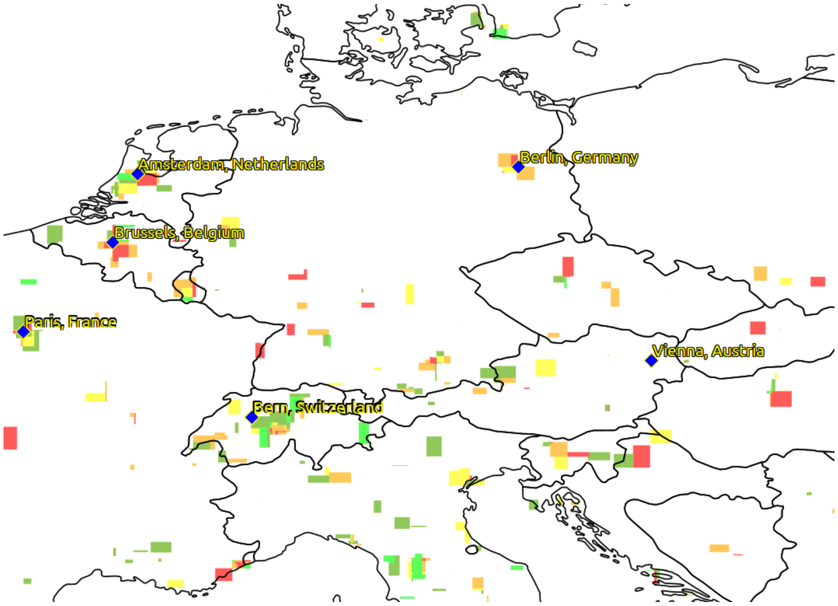


Fig. 8. Europe - geospatial opinion visualization

Human factors in computing systems, CHI EA '09, pages 3859–3864, New York, NY, USA, 2009. ACM.

12. R. Jianu and D. Laidlaw. Visualizing gene co-expression as google maps. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. Chung, R. Hammound, M. Hussain, T. Kar-Han, R. Crawfis, D. Thalmann, D. Kao, and L. Avila, editors, *Advances in Visual Computing*, volume 6455 of *Lecture Notes in Computer Science*, pages 494–503. Springer Berlin / Heidelberg, 2010.
13. D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 1 edition, 2000.
14. J. L. Leidner. Toponym resolution in text: annotation, evaluation and applications of spatial grounding. *SIGIR Forum*, 41:124–126, December 2007.
15. M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *International Conference on Data Engineering*, pages 201–212, 2010.
16. M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Steward: architecture of a spatio-textual search engine. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, GIS '07, pages 25:1–25:8, New York, NY, USA, 2007. ACM.
17. K. S. McCurley. Geospatial mapping and navigation of the web. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 221–229, New York, NY, USA, 2001. ACM.
18. B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series, 2010.

19. A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).
20. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.
21. R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. The design and implementation of spirit: a spatially aware search engine for information retrieval on the internet. *Int. J. Geogr. Inf. Sci.*, 21:717–745, January 2007.
22. G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 43–52, New York, NY, USA, 2010. ACM.
23. R. E. Roth, K. S. Ross, B. G. Finch, W. Luo, and A. M. MacEachren. A user-centered approach for designing and developing spatiotemporal crime analysis tools. Zurich, Switzerland, 14-17th September, 2010 2010. GIScience.
24. R. E. Schapire and Y. Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, 39(2/3):135–168, 2000.
25. N. Stokes, Y. Li, A. Moffat, and J. Rong. An empirical study of the effects of nlp components on geographic ir performance. *Int. J. Geogr. Inf. Sci.*, 22:247–264, January 2008.
26. B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 18:1–18:10, New York, NY, USA, 2008. ACM.
27. J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, pages 486–497, Mexico City, Mexico, 2005.
28. T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
29. J. Zhang, H. Shi, and Y. Zhang. Self-organizing map methodology and google maps services for geographical epidemiology mapping. *Digital Image Computing: Techniques and Applications*, 0:229–235, 2009.
30. W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, pages 354–362, New York, NY, USA, 2005. ACM.