Proceedings of the 23nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, M. Ali, M. Gertz, Y. Huang, M. Renz, and J. Sankaranarayanan, eds., Seattle, WA, November 2015.

Brands in NewsStand: Spatio-Temporal Browsing of Business News

Ahmed Abdelkader, Emily Hand, Hanan Samet *
Department of Computer Science
University of Maryland, College Park
College Park, MD 20740
{akader, ehand, hjs}@umiacs.umd.edu

ABSTRACT

The NewsStand system enables the use of a map query interface to retrieve news articles associated with the principal locations that they mention collected as a result of monitoring the output of over 10,000 RSS news feeds, made available within minutes of publication. NewsStand has been enhanced to allow using the map query interface to access other information associated with the articles such as photos and videos, as well as names of people and diseases mentioned in these articles. Here we report on our efforts to enhance NewsStand to display the names of brands and to the articles mentioning them. The challenges in identifying interesting brand mentions are discussed.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Business Intelligence

Keywords

Knowledge discovery, text mining, NewsStand, map query interface, geotagging

1. INTRODUCTION

Business news comprises a sizable portion of the daily document feed across the spectrum ranging from online sources to local newspapers. Besides the dedicated business section, various other news stories frequently have implications for businesses as well. Such news stories are significant in that they often influence the decisions of consumers, investors, and business owners which has the potential to quickly translate into profits and losses for the businesses mentioned. For this reason, business news has been used in stock market forecasts. As such, there is a real need to access the relevant news articles both thoroughly and efficiently.

To this end, we report on our experience in developing a set of features to answer this need within NewsStand - a spatio-temporal news browser [10, 14, 16, 19] that enables the news to be accessed by the locations that they mention achieved by using a map query interface. NewsStand crawls the web seeking news articles and tags each article with an associated location [8, 9, 11, 12] along with other attributes. It monitors the output of over 10,000 RSS news feeds, which are made available within minutes of publication, and automatically clusters articles into categories, taking into account geographic references and presents articles on an interactive map. NewsStand's user interface can be used to access other information associated with the articles such as photos and videos, as well as names of people and diseases that are mentioned in them. It is a natural extension to spatiotextual data of our earlier work on the SAND spatial browser [2, 15] and road networks [17]. It also works for tweets [4, 5, 18].

This paper relates our efforts to enhance NewsStand with the ability to recognize articles related to a subset of businesses across a variety of categories and introduces new interactive queries to provide a convenient way to explore this rich information across time and space. The demonstration allows users to query the system and browse the results on the map interface.

The rest of this paper is organized as follows. Section 2 discusses the issues and challenges that arise in our attempts to recognize relevant business articles. Section 3 details our approach. Section 4 describes an experimental evaluation. Section 5 demonstrates a prototype of the new features in our system. Finally, future work is discussed in Section 6.

2. RECOGNIZING RELEVANT ARTICLES

As described in the introduction, NewsStand automatically fetches news articles and labels their contents with various attributes that we simply refer to as *tags*. These tags allow specialized views on top of the map known as *layers*. In this framework, we seek a method to recognize when a business is mentioned to enable applying a special tag, i.e. `BUS` for business. Once the tag is applied, the user will be able to browse these articles in a dedicated *brand layer*. In addition, the map-query interface will allow the user to fur-

*Supported in part by the National Science Foundation under Grants IIS-12-19023, and IIS-13-20791.

SIGSPATIAL'15 November 03-06, 2015, Bellevue, WA, USA
ACM 978-1-4503-3967-4/15/11.
http://dx.doi.org/10.1145/2820783.2820795.

ther query this subset of articles using any combination of keywords, location and time.

More specifically, we need to recognize business entities in the body of an article, viewed as a sequence of sentences. Such an entity can then be tagged by creating a record referencing the article and specifying the substring and the tag applied to it. The substrings denoting a given business entity is what we refer to as a *brand*. As it turns out, recognizing relevant brands presents three particular challenges explained below.

The first challenge is that brands are not known beforehand. Furthermore, new brands are constantly being created whenever a new product is introduced or a new business is created. The second challenge is to disambiguate brand words that happen to be normal English words, e.g., Apple, Fidelity, Orange, United, etc. The third challenge is to discern the capacity in which the business entity is treated in the text. If someone *googles* something or shares it on Facebook or Twitter, this typically has little to do with these businesses. Another example is an article about a car accident involving a Ford Focus. We desire some degree of control on the type of contexts we include in the brand layer.

One way to go about this is to design a classifier that for each word or sequence of words determines two flags: a brand flag and a business context flag. When the brand flag is `YES`, then this means that we have a business entity. On the other hand, when the context flag is `YES`, it means the context is relevant for the purposes of the brands layer. The combination of the two flags gives rise to four different classes denoted as ordered pairs of the form (brand flag, context flag). Given such classifier, the `BUS` label would be applied to entities assigned to the `(YES, YES)` class.

For the prototype we present in this demo, we simplify the problem as stated above by fixing a list of brands with which we work. 1000 brands were obtained from the web and compiled into a dictionary file for the system to use. Using the dictionary, we concern ourselves with classifying only the occurrences of brand words in this dictionary. This means we decide beforehand the specific words to look for and only attempt to filter out irrelevant instances that violate either the brand or the context requirement.

3. APPROACH

We experimented with four approaches to the simplified version of the problem: a baseline approach, two rule-based approaches and a supervised learning approach. For the baseline, any occurrence of a word from the dictionary is assigned a `(YES, YES)` label, i.e., string matching.

The first rule-based approach searches for words in local context windows around brand instances that would indicate a true mention or a false mention. For each brand instance of the same brand in a document, we extracted the sentence around the brand as the context. So if Google was mentioned twice in the document, the context for Google in that document would be the two sentences containing the mention of Google. Each document may contain instances of multiple brands, and so we do this for each brand in each document. To find correct brand mentions, we look at the previous word and the word following the brand instance, if either are capitalized, then it is likely that the brand instance is a part of a larger brand name or a title (Orange in "Orange County"). Counting the number of times the previous word was either 'of' or 'in' helps to eliminate locations ("Mr. Smith of Kansas" with brand name Kansas). Using NLTK [1], we compute the part of speech tags for the contexts, and look at the part-of-speach (POS) of the brand instance in each context. If we find that the brand is an adjective, we say that it is a false brand mention ("Strong improvements today" for brand name Strong). To improve the accuracy of context mentions, we use some of the same features, as a brand instance that is a false brand mention usually suggests false context mention. In addition to these features, a ratio of the number of instances of a brand to the length of the document can indicate whether the article is about a brand or if it is just being mentioned in passing.

The second rule-based approach makes use of StanfordNER. StanfordNER is an open source library for named entity recognition that comes with a set of pre-trained classifiers encoded as binary files. In particular, StanfordNER [3] recognizes entities of the following types: `Person`, `Location` and `Organization`. Our problem can be stated as a refinement of the entity recognition problem that should take care of brand word disambiguation. With that, we only consider entities recognized by StanfordNER as `Organization`. If in addition the entity is one of the brands we included in the dictionary, we assign to it the `(YES, YES)` label. Further restricting the context could also be handled by topic modeling approaches.

For the supervised learning approach, we build a classifier that maps (brand, contexts) to (brand flag, context flag) for a given document. We have three levels of features for this task: brand, context and document. Motivated by the improvements achieved by the rule-based approach, we again make use of shallow text features [6], e.g., number of mentions, average length of sentencees and brand words, number of words, words with capitalization. In addition, we use two features to capture the prominence of mentions on the context and document level. A brand is marked as prominent within its context if it appears in the first half of the sentence, which loosely corresponds to POS approaches. Similarly, each context sentence may or may not be considered prominent in the document, which could be taken to suggest the brand is or is not relevant to the entire document, depending on the location of the sentence in the document. We simply map the position of the sentence to one of 4 bins corresponding to splitting the document into four parts. A better approach could use paragraph counts. We used Vowpal Wabbit [7] to learn a classifier on combinations of these features and selected a subset that gives the best accuracy.

4. EXPERIMENTAL EVALUATION

NewsStand provided us with a large set of data to work with. To prepare a data set for testing each of the approaches we described in Section 3, we started by implementing the baseline approach into NewsStand such that it returns a set of documents with their associated text and the brands mentioned in each document.

4.1 Data

We created a data set of 940 documents spanning a total of 297 distinct brands. For each of the brand instances, we manually assigned two flags to each brand instance: one for brand correctness and one for context correctness, as discussed in Section 2. We were concerned with improving the results for the brands NewsStand was already finding. Since the baseline approach easily achieves 100% recall for

the simplified version of the problem, we wanted to reduce the number of false positives. So for each document, we only sought to remove bad brand references returned from NewsStand.

4.2 Results

To simplify the presentation, we focus on the final decision of whether or not a given entity should be included in the brands layer. Among all the four classes each entity may be assigned to, only the `(YES, YES)` class is included. As such, we simply refer to this class as `TRUE` and refer the union of the remaining three classes as `FALSE`. Our data set contains 1190 instances, as each article can have more than one occurrence of the same entity. Of these instances, 365 are `TRUE` and 825 are `FALSE`. For the supervised learning method, we split the data set into a training set with 255 `TRUE` and 563 `FALSE` instances, and a test set of 110 `TRUE` and 262 `FALSE` instances.

Classifier	Precision	Recall	F-measure
BL	0.307	1.000	0.469
R1	0.670	0.501	0.573
R2	0.474	0.605	0.532
VW	0.703	0.818	0.756

Table 1: Precision and Recall for all four classifiers described in Section 3.

Table 4.2 summarizes the results of our experiment. The baseline method (BL) which always returns `TRUE` for all instances, since they match the strings in the dictionary of brands, trivially achieves 100% recall but only 30.7% precision. The first rule-based approach (R1), which uses POS tags and shallow features, achieves a significant improvement in precision bringing it to 67% but only at the expense of recall missing 49.9% of the `TRUE` instances. The second rule-based approach (R2), which uses StanfordNER to filter the instances, reverses the situation with higher recall at 60.5% with precision falling down to 47.4%. Finally, the supervised learning approach (VW) achieves the highest precision at 70.3% with a superior recall of 81.8% on the test set. On the training set, VW achieves 84.2% precision and 96% recall. It is interesting to note that this supervised learning approach only uses shallow text features [6]. The success of the supervised learning approach encourages us to pursue a more carefully tailored machine learning model.

5. DEMONSTRATION

In this section, we describe the prototype of the new features we implemented in NewsStand. We do this through a step-by-step tutorial that explains how to activate the brands layer and use it to browse news articles recognized as relevant for businesses in addition to performing basic queries like keyword search. This builds on top of the spatio-temporal filtering capabilities available in NewsStand, which are accessible to the brands layer.

In order to get started, the user needs to point their browser to: `http://newsstand.umiacs.umd.edu/`. While most features of NewsStand work on mobile phones as well, the scenarios below have only been tested on desktops.

5.1 Brands Layer

To activate the brands layer, the user clicks the "Settings" button at the bottom right corner of the screen and chooses "Brands" from the "Layers" list. Once the map view is updated, the user is presented a set of labels with a subset of the brands found in recent articles. The user can zoom in on a specific location to examine the articles there at a finer scale. As NewsStand clusters articles around the same location, each label actually represents a small set of articles that the user can access by clicking the label.

Figure 1 illustrates news browsing on the map interface. Both the snippets of the article mentioning both the brand and the associated location are shown in addition to a *mini-map* [13] that highlights other locations where the brand in question has been mentioned.

5.2 Map Mode Queries

Map mode is the default mode of browsing in NewsStand. If the user cannot immediately find the brands they are interested in, then they can use "Keyword Search" to filter the results. The user is also able to control how many results to display by adjusting the slider to the right.

5.3 Time Mode Queries

To explore the temporal span of a given query, the user can switch to the "Time Mode". In this mode of browsing, each label corresponds to the entire history of articles referencing that location. This history can be accessed through the minimap which includes a time slider at its top when "Time Mode" is activated. By adjusting the time window to query on this slider, the user can then advance the query in time by starting at the far left and dragging the window along the slider. The minimap displays a heat map overlay on top of the world map that shows the concentration of articles returned by the query on different locations around the globe. This allows the user to see the progression of events relevant to the query across both time and space.

6. DISCUSSION AND FUTURE WORK

In this paper, we introduced the problem of recognizing news articles which are relevant from a business perspective. We outlined the challenges involved in addressing this problem and demonstrated a set of interesting features that can be developed if such information is provided.

One shortcoming of our work is the result of a relatively small data set which was mainly limited by our ability to manually label only so many instances to obtain training and testing examples. Consequently, we avoided using context words due to the limited number of labeled articles and the diversity of contexts for brands from different industries and the type of news story in the article.

This demo only scratches the surface as we have not yet developed a rigorous solution to the underlying tagging problem. The simplified version of the problem we used for this prototype ignores crucial linguistic aspects since it only focused on sentences mentioning a given brand explicitly. In addition, we anticipate further improvements by using article metadata which includes: publication date, URL, topic, description and cluster information.

We are actively developing the map interface [13] as we firmly believe it plays a key role in providing efficient access to data, especially on mobile devices. Finally, we envision a more general framework to generate specialized layers for

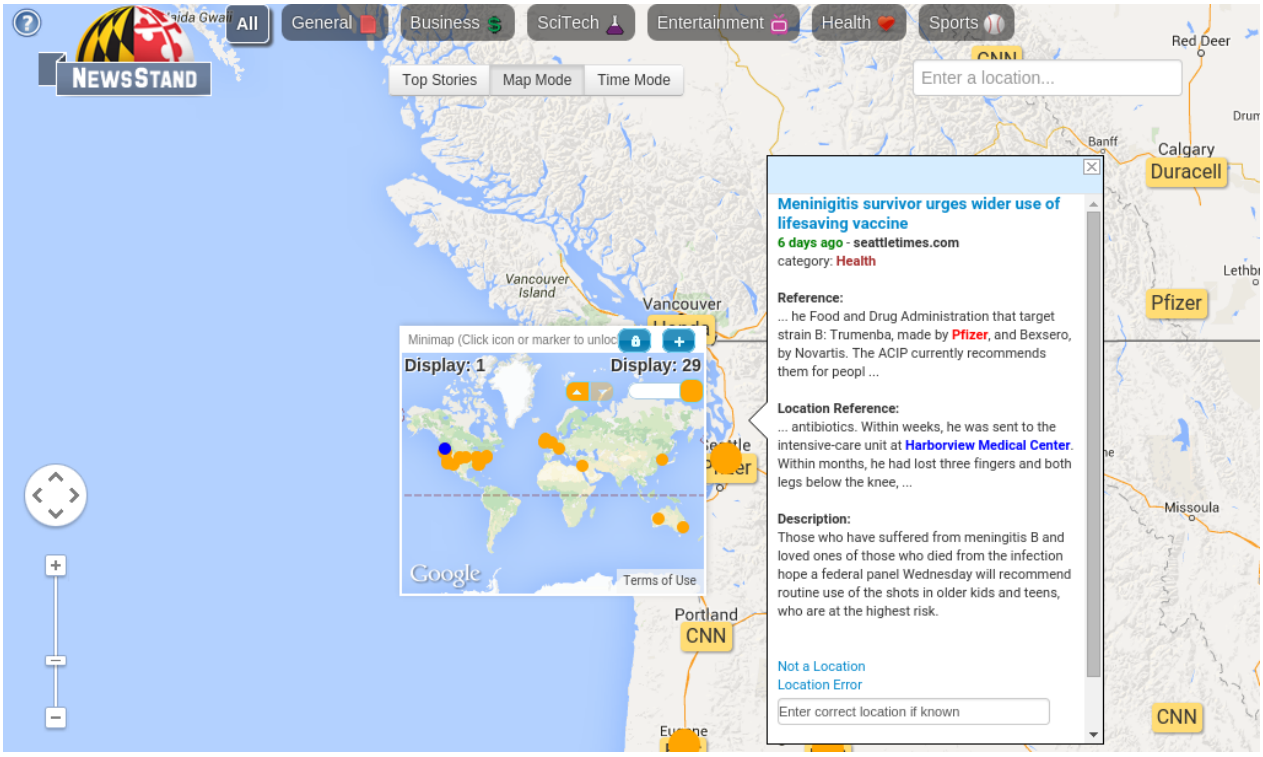

Figure 1: Example of "Map Mode" browsing in the brands layer.

any given category. For instance, the user may provide a set of keywords, queries or articles, and expect recurring results stemming from the examples they provided.

7. REFERENCES

[1] S. Bird. Nltk: the natural language toolkit. In *Proc. COLING/ACL on Interactive Presentation Sessions*, pages 69–72, Jul. 2006.

[2] C. Esperança and H. Samet. Experience with SAND/Tcl: a scripting tool for spatial databases. *JVLC*, 13(2):229–255, Apr. 2002.

[3] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Jun. 2005.

[4] N. Gramsky and H. Samet. Seeder finder - identifying additional needles in the Twitter haystack. pages 44–53, Orlando, FL, Nov. 2013.

[5] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using Twitter. In *LBSN*, pages 25–32, Chicago, Nov. 2011.

[6] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Proc Third ACM Intl. conf. on Web search and Data Mining*, pages 441–450, Feb. 2010.

[7] J. Langford, L. Li, and A. Strehl. Vowpal wabbit. *URL https://github.com/JohnLangford/vowpal_wabbit/wiki*, 2011.

[8] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *SIGIR'11*, pages 843–852, Beijing, China, Jul. 2011.

[9] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *SIGIR'12*, pages 731–740, Portland, OR, Aug. 2012.

[10] M. D. Lieberman and H. Samet. Supporting rapid processing and interactive map-based exploration of streaming news. In *GIS*, pages 179–188, Redondo Beach, CA, Nov. 2012.

[11] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *GIR*, pages 6:1–6:8, Zurich, Switzerland, Feb. 2010.

[12] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE*, pages 201–212, Long Beach, CA, Mar. 2010.

[13] H. Samet. Using minimaps to enable toponym resolution with an effective 100% rate of recall. In *GIR*, pages 9:1–9:8, Dallas, TX, Nov. 2014.

[14] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. Porting a web-based mapping application to a smartphone app. In *ACM GIS*, pages 525–528, Chicago, Nov. 2011.

[15] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *CACM*, 46(1):63–66, Jan. 2003.

[16] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *CACM*, 57(10):64–77, Sep. 2014.

[17] J. Sankaranarayanan, H. Samet, and H. Alborzi. Path oracles for spatial networks. *PVLDB*, 2(1):1210–1221, Aug. 2009.

[18] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51, Seattle, WA, Nov. 2009.

[19] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *GIS*, pages 144–153, Irvine,CA, Nov. 2008.