International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Understanding Emerging Spatial Entities Through KB Harvesting

Gaddipathi Bharathi[*1], Rachapudi Venkata Ratna Sirisha[2]
[*1]Assistent Professor, Department of MCA, St. Mary's Group of Institutions, Guntur, Andhra Pradesh, India
[2]PG Students, Department of MCA, St. Mary's Group of Institutions, Guntur, Andhra Pradesh, India

ABSTRACT

New entities are being created daily. Though the novelty of these entities naturally attracts mentions, due to lack of prior knowledge, it is more challenging to collect knowledge about such entities than pre-existing entities, whose KBs are comprehensively annotated through LBSNs and EBSNs. In this paper, we focus on knowledge harvesting for emerging spatial entities (ESEs), such as new businesses and venues, assuming we have only a list of ESE names. Existing techniques for knowledge base (KB) harvesting are primarily associated with information extraction from textual corpora. In contrast, we propose a multimodal method for event detection based on the complementary interaction of image, text, and user information between multi-source platforms, namely Flickr and Twitter. We empirically validate our harvesting approaches improve the quality of KB with enriched place and event knowledge.

Keywords : KB harvesting, Location based Social Networks, Event based Social Networks, Emerging Entity, Event Detection

I. INTRODUCTION

Location-based Social Network (LBSN like Foursquare and Google+ Local is a location-based knowledge base (namely, place KB) providing opportunity for users to share their offline experiences online. However, LBSN services suffer from emerging spatial entities (ESEs) such as new business and venue, reported to be 1% of place KBs. Given that their KB pages have little to no meaningful description until experts or non-expert volunteers annotate with offline check-ins (*e.g.*, comprehensive KB pages like Wikipedia articles take on average 133 days to be documented for ESEs users are restricted from obtaining substantial Information on any ESE. The role of the place KB is even more important recently, With the advent of Event-based Social Network (EBSN) such as Meet up and Eventbrite, integrating users' event experiences on a spatial level. However, as a real-world event is generally recognized as a more extended tuple *(location, topic, time)* the sparseness problem of place KBs may be multiplied, requiring additional event information per place. To overcome such problems, it is critical to identify immediate and automatic information harvesting techniques for LBSN and EBSN. The key claim of this paper is that we can support LBSN and EBSN by harvesting place and event knowledge from multimodal social media. Our goal, to illustrate by an example, is to harvest Flickr and Twitter postings attracted by US presidential debate between Clinton and Trump at Washington University in St. Louis, describing the venue, image, time, and topics, for future uses in LBSN and EBSN. We argue why these rather unconventional sources of KB harvesting are effective as follows. First, Flickr provides co-occurring words with high semantic similarity to Wikipedia.

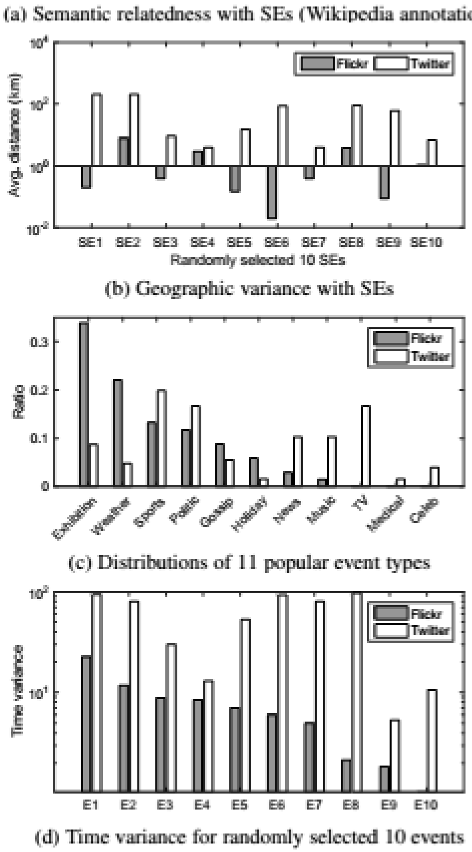


Figure 1 (a) randomly selects 10 SEs (with Wikipedia annotation) and identifies top-10 most frequently cooccurring context words per SE from Flickr and Twitter, from which, the semantic similarity scores [8] with SEs are much higher in Flickr than in Twitter context. Such co-occurrence helps eliminating ambiguity of whether to link 'Washington University' to the one in St. Louis or Seattle. Figure 1 (b) also shows that the precision of locational signal of Flickr is extremely high. Second, Flickr alone is not sufficient as a popular discussion topic on one modality of social media platform can rarely be mentioned in another modality. For example, political discussions are rarely captured in Flickr but frequent in Twitter. More generally, Figure 1 (c) shows such differences of event types [9] frequently mentioned in Flickr and Twitter, showing complementary nature. Third, despite promise, these sources have limitations and varying characteristics, complicating a joint modeling. Especially, unlike Flickr photos taken on time when the debate event occurs, Twitter posts allows users to post before, during, or after the event, which makes it challenging to capture the burstiness of event mentions. Figure 1 (d) randomly selects 10 events mentioned in both Flickr and Twitter, where photos are temporally closer to their referent events (7 hours from the photographed time) than tweets (68 hours from the uploaded time) on average (not counting retweets).

II. RELEATED WORK

An existing solution for improving recall is clustering duplicate or near-duplicate photos to collect missing or synonymous tags (which we use as a baseline in this work). We propose a more systematic way to tackle C1 and C2 and achieve significantly higher recall over baselines. This recall gain is critical in many applications– F1 score of SE type categorization using photos is improved by 7.7% In particular; we propose a systematic approach for photo population by aggregating multimodal signals.

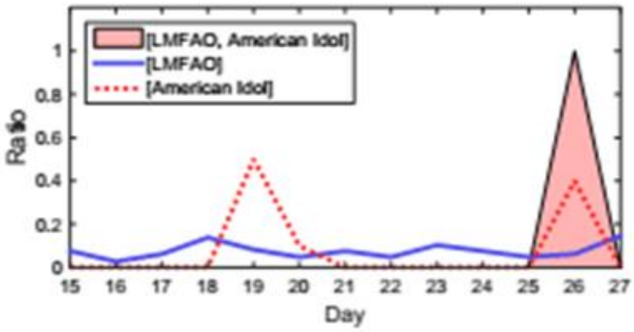
The sparseness problem of place KBs may be multiplied, requiring additional event information per place.

Photo Harvesting

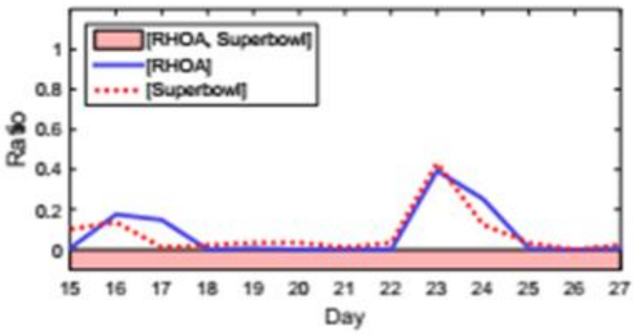
On word level, ImageNet [45] and ImageKB [10] are public large-scale labeled photo collection based on WordNet. On entity level, Taneva et al. [46] is closest to our work. A key difference with this work is that we assume minimal information from KB, *e.g.*, empty (or near empty) KB pages, to support emerging entities. We use discovering [16], [17], [47] for harvesting baseline REC and SYN. However, they are less robust being a simple aggregation.

Photo Tag Expansion. Tag annotation work has been applied to mine tags (for tagless photos that are more than 50%) [15], [48], by exploiting tag co-occurrence. More specialized approaches use geo-spatial neighbor [49], [50] and visual neighbor [51], [52].

Synonym Expansion. Finding entity synonyms is, as shown in Section 3.1.2, important for KB harvesting, especially for emerg-ing entities with ambiguous names [20]. Entity profile, redirection, and cross-lingual links of Wikipedia have been actively leveraged. For emerging entities with no such resource, Cheng et al. [16] and Chakrabarti et al. [17] leverage the term co-occurrence among user-generated contents to find entity synonyms.



(a) Related keywords with not similar bursts



(b) Not related keywords with similar bursts

Figure 2. Different event grounding results between MEG and EDCoW

III. PROPOSAL WORK

KB Harvesting for LBSN:

A naive way to harvest photos on KB is querying SE names (*e.g.*, 'Washington University') using location as an additional feature to disambiguate This method is highly precise but suffers from low recall. In contrast, we propose high-recall 96approaches, tackling the following two causes of low recall:

• Tag Sparsity: Web-photos may be poorly tagged or often have no "identifying" tag such as SE name.

• Vocabulary Mismatch: People tend to use various "synonymous" tags, which refer to the same SE but cannot be matched with its queried official SE name.

KB Harvesting for EBSN.:

This paper is an extended version of a conference paper that appeared as The key addition of this journal version is the interlinking between offline check-ins (*i.e.*, place knowledge) and event experiences in social media in higher accuracy and coverage. Most of existing event detection approaches are restricted to leveraging a single source of social media information. Instead of, for example, Chen and Roy and Ritter et al. Leverage only Flicker and Twitter, respectively, for event detection and annotation. Meanwhile, it is non-trivial to extend the Flicker- or Twitter-based techniques to mutually support the other modality platforms. One of the possible solutions would be aggregating detection results from each social media platform in event-level. However, this naive aggregation approach does not consider the complementary nature of the two sources in feature level, which we claim to significantly improve accuracy of event detection and annotation.

Harvesting Photos from Twitter:

We introduce the framework of our photo harvesting system. First, we collect photos and their metadata (user ID, image, location (*lat* and *lon*), tags (only if exists)), which are generally available in most photo-sharing sites such as Flickr. These photos are both geographically and visually grouped into duplicate/near-duplicate photo clusters

Textual Signal Estimation.

Given a photo cluster p, we compute $Pr(ne|p)$ which represents how relevant p is to ne. Intuitively, the probability is estimated by matching an aggregated tag set of p with ne. Such approach is likely to exclude photo clusters not annotated with SE names. To loosen it, a photo cluster p can be matched with an SE e, even though it is not annotated with its name, if p can be matched with another photo cluster pi with such annotation.

Image Features:

To quantify the image relevance between a tag and an SE, we extend the intuition of TFIDF. However, our unique contribution is to define a "document" as an estimated set *Pe* (*e.g.*, by a query *ne*) of all photos on the same SE *e*. Using this document, the following frequency features convey the synonym evidence in both TF and IDF.

Extracting Structured Knowledge from Photos

Wikipedia, a typical KB, has textual descriptions and metadata info box populated by human annotators. Though Flicker tags lack these rich descriptions, here, we discuss how to enrich an ESE tuple *⟨Ne,Pe⟩* with tags and metadata from Flickr photos, specifically to augment an ESE tuple *⟨Ne,Pe⟩* into a more extended ESE

tuple *⟨Ne,Pe,De,Ye,Le⟩* where *De* is context words, *Ye* is SE type, and *Le* is geographic distribution. We present the estimation of these three elements in detail.

IV. CONCLUSION

This paper studied the problem of KB harvesting for ESEs. Our proposed approach improves the quality of KB harvested by leveraging multimodal signals with respect to place and event. We demonstrate, for the first time, that multimodal interactions between different social media streams are critical in improving the quality of place KBs and capturing the temporal burstiness of event mentions.

V. REFERENCES

[1]. S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks." ICWSM, 2011.

[2]. E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in SIGKDD, 2011.

[3]. Z. Wu, Y. Song, and C. L. Giles, "Exploring multiple feature spaces for novel entity discovery," in AAAI, 2016.

[4]. K. Feng, G. Cong, S. S. Bhowmick, and S. Ma, "In search of influential event organizers in online social networks," in SIGMOD, 2014.

[5]. X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based social networks: linking the online and offline social worlds," in SIGKDD, 2012.

[6]. S. Unankard, X. Li, and M. A. Sharaf, "Emerging event detection in social networks with location sensitivity," World Wide Web, vol. 18, no. 5, pp. 1393–1417, 2015.

[7]. J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," 1998.

[8]. G. Tsatsaronis, I. Varlamis, M. Vazirgiannis, and K. Nørvåg, "Omiotis: A thesaurus-based measure of text relatedness," in Machine Learning and Knowledge Discovery in Databases, 2009.

[9]. A. Ritter, O. Etzioni, S. Clark et al., "Open domain event extraction from twitter," in KDD, 2012.

[10]. X.-J. Wang, Z. Xu, L. Zhang, C. Liu, and Y. Rui, "Towards indexing representative images on the web," in MM, 2012.

[11]. H. S. Packer, J. S. Hare, S. Samangooei, and P. Lewis, "Semantically tagging images of landmarks," in KECSM, 2012.

[12]. R. Srihari, C. Niu, and W. Li, "A hybrid approach for named entity and sub-type tagging," in ACL, 2000.

[13]. D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Map-ping the world's photos," in WWW, 2009.

[14]. Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven, "Tour the world: building a web-scale landmark recognition engine," in CVPR, 2009.

[15]. F. Belém, E. Martins, T. Pontes, J. Almeida, and M. Gonçalves, "Associa-tive tag recommendation exploiting multiple textual features," in SIGIR, 2011.

[16]. T. Cheng, H. W. Lauw, and S. Paparizos, "Entity synonyms for structured web search," In TKDE, 2012.

[17]. K. Chakrabarti, S. Chaudhuri, T. Cheng, and D. Xin, "A framework for robust discovery of entity synonyms," in SIGKDD, 2012