Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 525–530

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International Workshop on Data Mining in IoT Systems (DaMIS 2017)

Spatio-Temporal Contextualization of Queries for Microtexts in Social Media: Mathematical Modeling

Jae-Hong Park[a], O-Joun Lee[a], Joo-Man Han[a], Eon-Ji Lee[a], Jason J. Jung[a,], Luca Carratore[b], Francesco Piccialli[b]

[a]Chung-Ang University
84, Heukseok-ro, Dongjak-gu, Seoul, Republic of Korea 06974
[b]University of Naples "Federico II"
Corso Umberto I, 40, Napoli, Italy 80138

Abstract

In this paper, we present our ongoing project on query contextualization by integrating all possible IoT-based data sources. Most importantly, mobile users are regarded as the IoT sensors which can be the textual data sources with spatio-temporal contexts. Given a large amount of text streams, it has been difficult for the traditional information retrieval systems to conduct the searching tasks. The goal of this work is *i*) to understand and process microtexts in social media (e.g., Twitter and Facebook), and *ii*) to reformulate the queries for searching for relevant microtexts in these social media.

Keywords: Query contextualization; Spatio-temporal contexts; Information fusion.

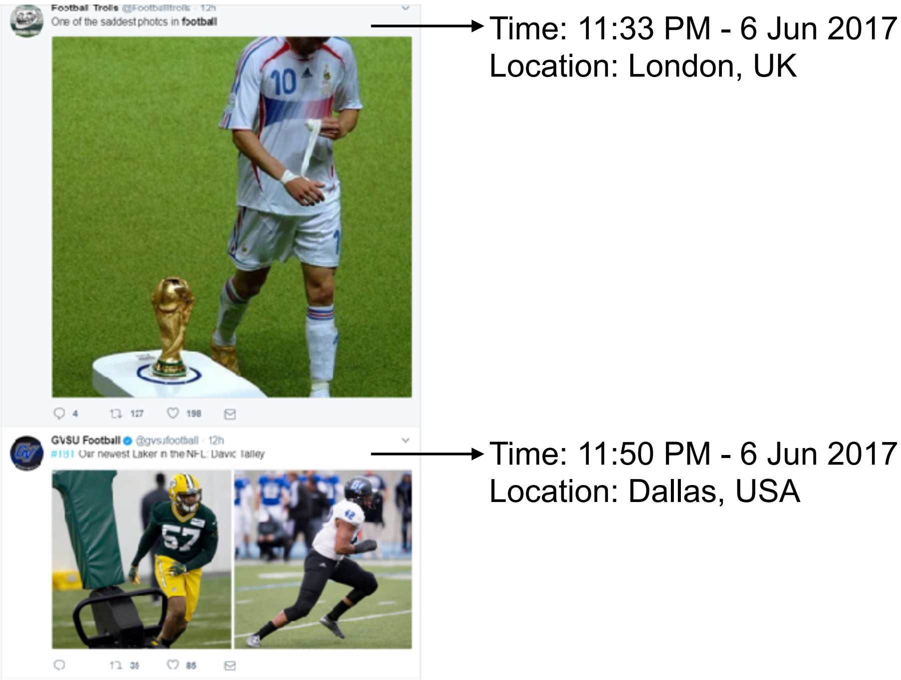
1. Introduction

Users with various mobile devices (e.g., smartphones and wearable devices) are regarded as the sensors from which text streams can be collected. The contents in the texts are related to personal stories as well as our society. More importantly, since the users are actively participating into popular social networking services (e.g., Twitter and FaceBook), the texts can be annotated with several metadata including location and time. Thus, somehow, we can figure out when and where a certain social event happened.

For example, if we search for "Football", we get results which include both of "football" and "soccer", as shown in Fig. 1. It is caused by that British peoples call "soccer" as "football." Therefore we should consider spatial contexts, if we want to find some social events like NFL (National Football League) of United States or The Premier League of United Kingdom.

* Corresponding author. Tel.: +82-2-820-5136 ; fax: +82-2-822-5301.
E-mail address: j3ung@cau.ac.kr

10.1016/j.procs.2017.08.317



Football Trolls @Footballtrolls · 12h
One of the saddest photcs in **football**

4 127 198

GVSU Football @gvsufootball · 12h
#181 Our newest Laker in the NFL: David Talley

35 85

Time: 11:33 PM - 6 Jun 2017
Location: London, UK

Time: 11:50 PM - 6 Jun 2017
Location: Dallas, USA

Fig. 1. An Example of Ambiguous Queries

In conventional search engines, this issue can be solved with relatively simple methods, since ordinary web documents are long enough to consider contexts of how words are used. To understand why it is different in the social media, we need to understand the main characteristics of the text streams collected from these SNS.

- A large amount of text streams
- Each of the texts is short.

These characteristics cause serious problems for social sensing and retrieving microtexts from social media. Queries inserted by users are usually short and ambiguous. In users' contexts, the queries might have relatively clear meanings. As shown in Fig. 1, a meaning of the word "Football" is definite with considering the users' spatial contexts. Also, we can consider various acronyms which are used as separate meanings according to domains (e.g., AI).

In conventional IR services including search engines, we can deal with these issues by using conventional methods which are efficient to discover semantics of the words (e.g., TF-IDF, LSA, LDA, and so on). However, in microtexts, these methods do not work well because of sparsity and narrow width of word-document matrices.

It makes us require to utilize the contexts of the users what the microtexts abundantly include. In this study, we propose a method for query expansion with contextual information, and we call it as "Query Contextualization." It enables us to disambiguate the queries by sensing and rebuilding the users' contextual information. As an early stage of this research, we use only temporal and spatial information which are easily gathered from the social media, in this paper.

The proposed method mainly consists of three steps: (i) grouping the words based on trends of co-occurrence frequencies with the terms in queries, (ii) discovering spatio-temporal singular points of each group, and (iii) providing users the semantic groups of words with their singular points. It enables users to choose an appropriate set of words which is matched with their intention. Furthermore, the spatio-temporal singular points do not only guide users' selection, but also restrict search spaces and increase precision of the retrieval.

The rest of this paper is organized, as follows. In Sect. 2, we describe the raising problem and introduce related studies with this paper. In Sect. 3, we present the proposed method how we contextualize the queries and provide them interactively to the users, and then in Sect. 4 we conclude our work and present a future direction of this study.

2. Social Sensing by Crowdsourcing

Social sensing is to detect social trends or events. On the past, it was an area of social scientists and questionnaire surveys. However, in these days, it is conducted by computers based on information what users voluntarily provide through the web. 'Google Trends[1]' is one of its representative examples.

Social media are one of the most utilizable data sources for social sensing by crowdsourcing, since users spontaneously share a massive amount of microtexts through the social media. In here, users of the social media take a role which is a sensor[1].

To represent trends or events detected from the microtexts, an adequate retrieval method is required. However, as shown in Fig. 1, we have to deal with ambiguities of the queries to provide accurate information of social events to the users. For reducing the ambiguities, query expansion methods are widely studied in the IR domain.

To reformulate queries on searching microtexts, Pseudo-Relevance Feedback has been widely applied[2,3,4]. This method is built on a hypothesis that the most relevant document with an initial query includes the most adequate words to be used for expanding the initial query. However, in social media with overflowing microtexts, the hypothesis is not always suitable[5,6]. Informal, rhetorical, and metaphoric terms included in the microtexts make it much worse.

Furthermore the most relevant document with initial queries are depending on the contexts. It is not only for a case of the spatial contexts which is shown in Fig. 1, but also for temporal contexts. Let suppose that the initial query is an abbreviation 'AI'. 'AI' has various meanings including Artificial Intelligence, Avian Influenza, and so on. An actual meaning of this term in the initial query is up to users' intention. However, if there is avian influenza pandemic, a document related with avian influenza will be chosen as the most relevant one, no matter what users originally intend.

To improve this issue, Wang et al.[7] have proposed a query expansion methods based on a feedback concept model. The feedback concept model which is also suggested by the authors is built on knowledge and information derived from Probase[2]. It enables search engines to reflect the users' intention within a concept-layer, not only a term-layer. However, the authors overlooked that the simplest way how we recover the contextual information is extracting and utilizing them from microtexts. Also, since the users of social media voluntarily provide their contexts, it is relatively easy to extract contextual backgrounds of the microtexts.

3. Query Contextualization

Queries which are usually inserted by human beings contain ambiguities. It is caused by that our everyday languages include various synonyms, homonyms, and acronyms. This problem is going worse, as short as the queries are. Typical search engines which deal with web documents apply information retrieval (IR) methods including automated query expansion methods to improve this issue.

However, the microtexts from social media contain few words in each document. It makes search engines for the microtexts hard to adopt the existing IR methods. To solve this problem, we apply spatio-temporal contexts attached on the microtexts as geo-tags and time-tags to expand queries on retrieving the microtexts.

Therefore an objective of our proposed method can be defined as: interactive query expansion for microtexts with considering spatio-temporal features. Also, it can be formulated as

$$\mathbf{T}(q, \langle\tau, \zeta\rangle) = \{q'_1(\tau_1, \zeta_1), q'_2(\tau_2, \zeta_2), \cdots, q'_N(\tau_N, \zeta_N)\}. \tag{1}$$

where $\mathbf{T}$ is a transform function, q indicate an original query, q' denotes a transformed query, and τ and ζ respectively mean temporal and spatial contexts.

3.1. Spatio-temporal contexts in social media

To expand the queries based on the spatio-temporal contexts, we firstly have to discover semantic relevancies among the words according to spatio-temporal changes. Let suppose that our target query is Q and $q_a \in Q, a \in [1, A]$

[1] https://trends.google.com/trends/?hl=sw

[2] https://www.microsoft.com/en-us/research/project/probase/

denotes a term in $\mathcal{Q}$. For exposing spatio-temporal relevancies between the term q_a and other word in the microtexts $w_i \in \mathcal{W}, i \in [1, I]$, we are focusing on co-occurrence frequencies among them according to the spatio-temporal contexts.

When τ and ζ respectively represent particular temporal and spatial locations and $\mathcal{M}_\tau$ and $\mathcal{M}_\zeta$ respectively denote sets of microtexts generated on τ and ζ, we can respectively formulate total co-occurrence frequencies between q_a and w_i on τ and ζ as

$$f^{co(q_a)}(w_i, \tau) = \sum_{m_l \in \mathcal{M}_\tau} f^{co(q_a)}_{m_l}(w_i), \tag{2}$$

$$f^{co(q_a)}(w_i, \zeta) = \sum_{m_l \in \mathcal{M}_\zeta} f^{co(q_a)}_{m_l}(w_i), \tag{3}$$

where $m_l \in \mathcal{M}, l \in [1, L]$ means a microtext in an universal set of microtexts $\mathcal{M}$ and $f^{co(q_a)}_{m_l}(w_i)$ denotes a co-occurrence frequency between q_a and w_i within m_l.

Based on the total co-occurrence frequencies, we can represent temporal and spatial changes of the semantic relevancies among the words as vectors and matrices. It can be respectively formulated as

$$\vec{T}_{q_a,w_i} = \langle f^{co(q_a)}(w_i, \tau_1), \cdots, f^{co(q_a)}(w_i, \tau_N) \rangle, \tag{4}$$

$$S_{q_a,w_i} = \begin{bmatrix} f^{co(q_a)}(w_i, \zeta_{1,1}) & \cdots & f^{co(q_a)}(w_i, \zeta_{1,K}) \\ \vdots & \ddots & \vdots \\ f^{co(q_a)}(w_i, \zeta_{K,1}) & \cdots & f^{co(q_a)}(w_i, \zeta_{K,K}) \end{bmatrix} = \begin{bmatrix} f^{co(q_a)}(w_i, \langle x_1, y_1 \rangle) & \cdots & f^{co(q_a)}(w_i, \langle x_K, y_1 \rangle) \\ \vdots & \ddots & \vdots \\ f^{co(q_a)}(w_i, \langle x_1, y_K \rangle) & \cdots & f^{co(q_a)}(w_i, \langle x_K, y_K \rangle) \end{bmatrix}, \tag{5}$$

where $\tau_n \in \mathcal{T}, n \in [1, N]$ is a certain temporal location, $\zeta_{j,k} \in \mathcal{Z}, j, k \in [1, K]$ means a particular spatial location, and x_j and y_k respectively represent latitude and longitude. The temporal and spatial locations are discrete and continual, and also designated with a regular intervals which are user-defined.

With the representations of the temporal and spatial changes, we build groups of the words according to their semantic relevancies. To make groups, we apply the k-means clustering algorithm which is well-known. Also, we define a metric for measuring qualities of cluster models to determine the number of clusters in the cluster models. This metric is based on an intra-compactness among the words within each cluster and an inter-adjacency between the clusters. It can be formulated as

$$\mathbb{Q} = \sum_{c_b \in C} \left\{ \sum_{w_i \in c_b} D(w_i, \mu_b) \cdot \alpha - \sum_{w_j \notin c_b} D(w_j, \mu_b) \cdot (1 - \alpha) \right\} \tag{6}$$

where c_b indicates a cluster in an entire cluster model C, μ_b means a center of c_b, $D(w_i, \mu_b)$ denotes a distance between the two words w_i and μ_b. In here, a preceding term of Eq. 6 indicates the intra-compactness of c_b and a succeeding term means the inter-adjacency.

Also, the method how we measure the distances between two words are different from whether we deal with temporal or spatial contexts. For the temporal contexts, we apply angular distance and euclidean distance. A semantic distance between w_i and w_j with the temporal contexts can be formulated as

$$D_{\mathcal{T}}(w_i, w_j) = D^A(w_i, w_j) \cdot D^E(w_i, w_j), \tag{7}$$

where

$$D^A(w_i, w_j) = \frac{1}{\pi} \times 2 \cos^{-1}(S(w_i, w_j)), \tag{8}$$

$$S(w_i, w_j) = \cos\theta = \frac{\vec{T}_{q_a,w_i} \cdot \vec{T}_{q_a,w_j}}{\|\vec{T}_{q_a,w_i}\|_2 \times \|\vec{T}_{q_a,w_j}\|_2}, \tag{9}$$

$$D^E(w_i, w_j) = \|\vec{T}_{q_a,w_i} - \vec{T}_{q_a,w_j}\|_2. \tag{10}$$

In case of the spatial contexts, we use Frobenius distance. A semantic distance between w_i and w_j with the spatial contexts can be formulated as

$$D_{\mathcal{Z}}(w_i, w_j) = \|S_{q_a,w_i} - S_{q_a,w_i}\|_F. \tag{11}$$

Finally, we can have two cluster models which are considering the temporal and spatial contexts, respectively. Each cluster within the cluster models is a set of semantically related words and represents a concept what the term q_a is used for.

3.2. Event detection

In Sect. 3.1, we built clusters of words which are used with terms in the queries. The clusters in the built cluster models are corresponded with concepts which are included in the terms. Based on these concepts, we can elicit spatio-temporal changes of dominant concepts.

When f_{τ_n,c_b} and $f_{\zeta_{j,k},c_b}$ indicate total frequencies of words included in c_b, frequencies of the concepts on particular spatio-temporal locations can be respectively formulated as

$$\vec{F}_{\tau_n} = \langle f_{\tau_n,c_1}, \cdots, f_{\tau_n,c_B} \rangle, \tag{12}$$

$$\vec{F}_{\zeta_{j,k}} = \langle f_{\zeta_{j,k},c_1}, \cdots, f_{\zeta_{j,k},c_B} \rangle. \tag{13}$$

Through these representations, we can find out which concept is dominant on a certain spatio-temporal location. If f_{τ_n,c_b} has the highest value in $\vec{F}_{\tau_n}$, it means that c_b is a dominant concept on a time point τ_n, and vice versa.

When Q_{τ_n} and $Q_{\zeta_{j,k}}$ respectively indicate dominant concepts on certain temporal and spatial locations, we can recognize shifts of the dominant concepts based on an equality between $Q_{\tau_{n-1}}$ and Q_{τ_n}. If the equality is retained from $Q_{\tau_{n-3}}$ to Q_{τ_n}, it means a social event is happened within a time interval $[n-3, n]$, and q_a was used as a corresponding concept in the time interval.

However, in case of spatial locations, it is hard to make particular interval, since the space is a two-dimensional. Therefore we use DBSCAN (Density-Based Spatial Clustering of Applications with Noise)[9] algorithm to make spatial regions which have common dominant concepts. To utilize DBSCAN, we represent concept's dominant locations as a matrix. When $\mathcal{S}_{c_b}$ marks dominant spatial locations of c_b, it is formulated as

$$\mathcal{S}_{c_b} = \begin{bmatrix} \mathcal{I}_{\zeta_{1,1}} & \cdots & \mathcal{I}_{\zeta_{K,1}} \\ \vdots & \ddots & \vdots \\ \mathcal{I}_{\zeta_{1,K}} & \cdots & \mathcal{I}_{\zeta_{K,K}} \end{bmatrix}, \tag{14}$$

$$\mathcal{I}_{\zeta_{j,k}} = \begin{cases} 1 & \text{, if } c_b = Q_{\zeta_{j,k}} \\ 0 & \text{, otherwise} \end{cases}. \tag{15}$$

Based on $\mathcal{S}_{c_b}$, DBSCAN algorithm makes regions with spatial locations which $\mathcal{I}_{\zeta_{j,k}} = 1$. In here, the minimum number of locations within a critical distance for DBSCAN algorithm is defined as 3, since we are dealing with a 2-dimensional space. Also, the critical distance is defined as

$$D^C = (1-\beta) \min_{\forall \zeta_{j,k}, \zeta_{l,m}} \|\zeta_{j,k} - \zeta_{l,m}\|_2 + \beta \max_{\forall \zeta_{j,k}, \zeta_{l,m}} \|\zeta_{j,k} - \zeta_{l,m}\|_2. \tag{16}$$

Finally, we can discover social events which have spatio-temporal aspects as temporal intervals and spatial regions.

3.3. Example on Query Reformulation

In this section, we present results of the proposed query contextualization method with a problem which we presented in Fig. 1. As shown in Table. 1, if we search the microtexts with a term "Football", it can include two distinguishable meanings: soccer and American football. In focus of temporal contexts, we search with "Football" on a day when games of Premier league or Champions league are held, the term might mean soccer. In case of spatial contexts, if we retrieve with "Football" at American soil, the term should denote American football.

Table 1. An Example of the Spatio-Temporal Query Contextualization

q	Temporal Context	Spatial Context	q'
Football	13:33 PM 6th Sep 2016	London, UK	Soccer, Premier League, Arsenal
Football	16:24 PM 16th Feb 2017	London, UK	Soccer, Champions League, Premier League, Arsenal
Football	13:50 PM 6th Jan 2017	Dallas, USA	American Football, NFL, playoff, cowboys
Football	11:50 PM 6th Sep 2016	Dallas, USA	American Football, NFL, kickoff, cowboys, Giants

4. Discussion and Conclusion

In this study, we proposed the method to reformulate the queries with considering spatio-temporal contexts of microtexts. For searching the microtexts, reflecting contextual information from the microtexts are quite simple, but efficient approach. Also, the proposed method enable the users to select expended queries which are composed based on the social events built on the spatio-temporal contexts. It makes the search engines consider the users' intention easily.

However, this study has crucial limitation as a preliminary work that the proposed method is not yet properly evaluated. In our future works, we will focus on designing and conducting adequate evaluation procedures for the proposed method.

Acknowledgements

This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW (20170001000011001) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Resch, B.. *People as Sensors and Collective Sensing-Contextual Observations Complementing Geo-Sensor Network Measurements*. Lecture Notes in Geoinformation and Cartography. Springer, Berlin, Heidelberg. ISBN 978-3-642-34202-8; 2013, p. 391–406. doi:10.1007/978-3-642-34203-5_22.
2. Liang, F., Qiang, R., Yang, J.. Exploiting real-time information retrieval in the microblogosphere. In: Boughida, K.B., Howard, B., Nelson, M.L., de Sompel, H.V., Sølvberg, I., editors. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2012)*. ACM. ISBN 978-1-4503-1154-0; 2012, p. 267–276. doi:10.1145/2232817.2232867.
3. Lv, C., Qiang, R., Fan, F., Yang, J.. *Knowledge-Based Query Expansion in Real-Time Microblog Search*; vol. 9460 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-319-28939-7; 2015, p. 43–55. doi:10.1007/978-3-319-28940-3_4.
4. Zhai, C., Lafferty, J.D.. Model-based feedback in the language modeling approach to information retrieval. In: Paques, H., Liu, L., Grossman, D., editors. *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2001)*. ACM. ISBN 1-58113-436-3; 2001, p. 403–410. doi:10.1145/502585.502654.
5. Cao, G., Nie, J., Gao, J., Robertson, S.. Selecting good expansion terms for pseudo-relevance feedback. In: Myaeng, S., Oard, D.W., Sebastiani, F., Chua, T., Leong, M., editors. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008)*. ACM. ISBN 978-1-60558-164-4; 2008, p. 243–250. doi:10.1145/1390334.1390377.
6. Miyanishi, T., Seki, K., Uehara, K.. Improving pseudo-relevance feedback via tweet selection. In: He, Q., Iyengar, A., Nejdl, W., Pei, J., Rastogi, R., editors. *Proceedings of the 2013 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2013)*. ISBN 978-1-4503-2263-8; 2013, p. 439–448. doi:10.1145/2505515.2505701.
7. Wang, Y., Huang, H., Feng, C.. Query expansion based on a feedback concept model for microblog retrieval. In: Barrett, R., Cummings, R., Agichtein, E., Gabrilovich, E., editors. *Proceedings of the 26th International Conference on World Wide Web (WWW 2017)*. ACM. ISBN 978-1-4503-4913-0; 2017, p. 559–568. doi:10.1145/3038912.3052710.
8. Fonseca, B.M., Golgher, P.B., Pôssas, B., Ribeiro-Neto, B.A., Ziviani, N.. Concept-based interactive query expansion. In: Herzog, O., Schek, H., Fuhr, N., Chowdhury, A., Teiken, W., editors. *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2005)*. ACM. ISBN 1-59593-140-6; 2005, p. 696–703. doi:10.1145/1099554.1099726.
9. Ester, M., Kriegel, H., Sander, J., Xu, X.. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad, U.M., editors. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*. AAAI Press. ISBN 1-57735-004-9; 1996, p. 226–231. URL `http://www.aaai.org/Library/KDD/1996/kdd96-037.php`.