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.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.

*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.
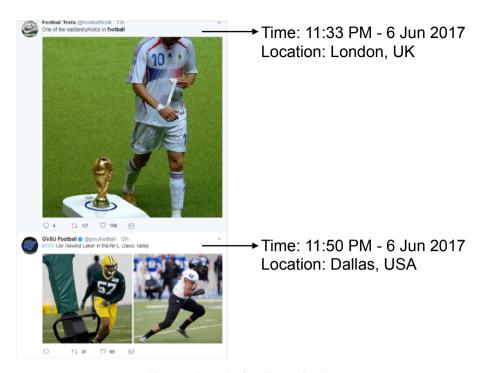  *E-mail address:* j3ung@cau.ac.kr

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.