Ahmet Yıldırım
Thesis Co-Supervisors: Suzan Üsküdarlı, PhD. And Assoc. Prof. Haluk Bingöl
TOPIC IDENTIFICATION WITHIN MICROBLOG POST COLLECTIONS

Abstract
This thesis aims to identify topics in collections of microblog posts, where topics correspond to a set of related topic elements. The first approach, BounTI, examines the use of Wikipedia -- well written cross-domain articles -- to capture topics within microblog posts that are messy, unstructured, and fragmented. The topic elements are identified based on their *tf.idf* scores, where the microblog post set is considered as a single document for *tf* computation. For *idf* computation, a public stream post set is used where each post is considered as a document. The *tf.idf* vectors of Wikipedia articles are computed, and the cosine similarity of the *tf.idf* vectors determine the topics. This approach was evaluated with more than 1 million tweets gathered during the 2012 US presidential election, resulting in a precision of 0.96 and F_1=1.

The second approach, S-BounTI, examines the generation of semantically structured topics, so that they can be further processed to yield more information. S-BounTI considers distinguishing elements of a post set as linked entities. Co-occurrence of two elements in the same post is considered as a relation. The related element sets which form topics are maximal cliques of the graph of elements and relations. To express topics, an ontology for microblog topics is introduced.
The topics can be utilized in conjunction with Linked Open Data (LOD). Over 1M posts during the 2016 U.S. presidential election debates, and other events such as the death of Carrie Fisher and the Dakota Access Pipeline demonstrations were considered for evaluation. Quantitative and qualitative observations are provided and example SPARQL queries and their results are presented to show the utilization of the topics. Both approaches gave promising results and are suitable for future research and development. S-BounTI has been found to represent related elements better then BounTI.

PUBLICATIONS

Journals

1. Yıldırım, A., Üsküdarlı, S., Özgür, A., Identifying topics in microblogs using Wikipedia. PloS one, 11(3), e0151885. 2016. (SCI-E)
2. Yıldırım, A., Üsküdarlı, S., Extracting Semantic Topics from Microblogs, Special issue on Linked Data for Information Extraction, Semantic Web Journal, *submitted*. (SCI-E)
3. Üsküdarlı, S., Yıldırım, A., Identifying Semantic Topics from Tweets: A Study of 2012 and 2016 U.S. Presidential Election Debates, Special issue on Computational Propaganda and Political Big Data, Big Data Journal, *submitted*. (SCI-E)

Conferences

1. Bingöl H., Habiboğlu M.G., Üsküdarlı S., Yıldırım A., Çalıkuş O., Sezgin C., Yelkenci S. An Operator Pprovided m-learning service: A Preliminary Report, IADIS International Conference, Mobile Learning, Porto, Portugal. 2010.
2. Yıldırım, A., Üsküdarlı, S., Semantic Tagging and Inference in Online Communities, in Proceedings of I-Semantics `08, Graz, Austria. 2008, pp. 174-177

Defense Jury Members

1. Assoc. Prof. Haluk Bingöl	Boğaziçi University
2. Suzan Üsküdarlı, PhD.	Boğaziçi University
3. Prof. Yağmur Denizhan	Boğaziçi University
4. Prof. Şule Gündüz Öğüdücü	İstanbul Technical University
5. Assoc. Prof. Arzucan Özgür	Boğaziçi University
6. Asst. Prof. Murat Can Ganiz	Marmara University

Defense Date: 07.06.2017