

Fribourg, 21. August 2019

Taxonomy Enrichment

MSc Thesis:

Work overview: FBTaxonomy (<https://fashionbrain-project.eu/fashion-taxonomy/>) is a fashion taxonomy developed in the context of the FashionBrain project (<https://fashionbrain-project.eu>). It contains 670 fashion items ordered hierarchically in 5 levels by merging multiple taxonomies coming from different online retailers, e.g., Zalando, Amazon, eBay, etc. The current version of the taxonomy is static and does not support adding new items.

The emergence of social media platforms offers the possibility to enrich FBTaxonomy with new fashion items. Existing taxonomy update techniques mainly classify topics by assigning them to a specific level of the taxonomy based on their similarities. Some of these techniques extract topics from textual data and represent them in a taxonomy, where every node represents a conceptual topic [1]. While others use a network embedding algorithm that reveals the similarities between nodes in a network [2]. These two approaches are not directly applicable in the fashion context where items do not necessarily represent topics.

The aim of this thesis is to implement a taxonomy enrichment technique that continuously updates FBTaxonomy. The outcome of this thesis should consist of the following points:

- represent fashion items as an embedding by learning a vector representation.
- compute the similarity between different fashion items embeddings.
- extract the similarity relations between items from short textual data (e.g., tweets) and long textual data (e.g., blogs).
- classify these fashion items in the FBTaxonomy by assigning each new item to the right level of the taxonomy.

Work tasks:

1. Compare existing embedding techniques to represent fashion items in FBTaxonomy as vectors.
2. Measure the similarity between the items embeddings using different techniques.
3. Investigate the possibility of measuring the confidence on assigning a fashion item to a certain level of the FBTaxonomy.
4. Evaluate the performance of the implemented techniques for taxonomy enrichment.
5. Write a thesis that describes the implemented techniques and the result of the experiments.
6. Make a presentation of 20 minutes.

References:

1. Ma, J., Cui, P., Wang, X. and Zhu, W. *Hierarchical taxonomy aware network embedding*. SIGKDD , 18.
2. Zhang, C., Tao, F., Chen, X., Shen, J., Jiang, M., Sadler, B., Vanni, M. and Han, J., 2018, July. *Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering*. SIGKDD, 18.
3. The code of Taxogen is available here <https://github.com/franticnerd/taxogen>.
4. The code of Nethiex is available here <http://nrl.thumedia.com/NetHiex-nCRP>

Starting date of thesis: TBD

Ending date of thesis: TBD

University of Fribourg
Department of Informatics
eXascale Infolab

Mourad Khayati, Ines Arous