Web Science and Engineering 2015

Homework 7

Eric Camellini
4494164
e.camellini@student.tudelft.nl

1. USER MODELING PAPER

The selected paper is [3].

The focus of that work is to propose a novel approach to re-rank tweets in the Twitter users' timeline, with the aim to show them ordered by relevance with respect to the users' interests. The user interest profile is derived from his/her tweets, represented as sets of concepts from Wikipedia, together with information about interactions with other users. The research questions are:

- What is the best way to represent tweets using Wikipedia concepts and data about involved users (users who post the tweet, and users who are mentioned or replied to)? What is the best way to build a user profile using the resulting tweet representation?
- Is it possible to exploit Wikipedia inter-link information in order to extend a user profile based on Wikipedia concepts?
- Does this Wikipedia-based user modeling outperform the TF-IDF modeling when used to re-rank tweets in the users' timeline?

Every tweet is represented using a weighted Wikipedia concept vector (using Explicit Semantic Analysis as described in [2]) and information on the involved users. The user profile is built from his/her tweets by adding up all the concepts and corresponding weights. It is then enriched with a user affinity vector (built by counting the number of tweets involving reply, retweet, or mention between two users) and with new concepts extracted through a Markov random walk on the Wikipedia graph (built using inter-links between Wikipedia pages). The timeline tweet ranking is based on the cosine similarity between user profile and tweets in the timeline. Accordingly to the recall-at-k, precision-at-k, and average hit-rank evaluation metrics this system outperforms the the TFIDF model described in [1].

2. LEARNER MODELING

Novel, non-trivial, features for modelling learners (i.e. users of a system for education or learning):

- *Average grade by field*: in order to model a learner it can be useful to consider his/her grades obtained in past courses by field of interest. This results in a set of features: one for every considered field. In this way it could be possible to recommend courses to the learner in fields where he/she performs good, or to rank the recommended courses by field (e.g. predicting the learner's expected performance).
- *Average time to drop*: this feature considers the average time that the learner spends on a course before dropping it (it could be also split by field as the previous feature). It can be used to recommend to the learner courses that have a maximum duration shorter than the feature value, so that he/she will more likely go though all the course without dropping it.
- *International teamwork experience*: this feature is the count of the hours that the user successfully spent working in international teams. Successfully means that in order to be counted in terms of hours, the work must be graded as passed. The hours of work cannot be really counted, but every assignment requires a certain amount of time (ideally, on average), so this is the hour count that could be considered.

  This feature could be useful in courses where teamwork is requested, so that it can be possible for users to find out how much time other users spent working with people from different countries before working with them, but also for the system in order to recommend courses that contains (or not) this kind of activities.

The described features can be used in MOOC (massive online open course) providers such as edX[1] or Coursera[2], where users follow courses, do assignments, and are eventually graded or evaluated in some way. For what concerns the last feature, as far as I know at the moment teamwork is not used by MOOC providers, but I think that it will be used in the future: people will need to work remotely with other people for some projects or assignments.

[1] urlhttps://www.edx.org/
[2] `https://www.coursera.org/`

The disadvantage of these features is that they suffer the cold-start problem: a new user does not have any of the data needed to compute them, so they become useless.

This problem derives from the fact that all these features need data that comes directly from the system that wants to model the learner (e.g. the MOOC provider can compute these three features using data that it personally owns and has collected).

A different way of reasoning could be to do a cross-system model of the user, gathering data from other systems: it could be possible, for example, to gather data from the Twitter profile of the user and use it to infer its interests, or from the LinkedIn profile (that is maybe more relevant when trying to infer academic and professional interests). This cross-system technique also solves the cold-start problem.

References

[1] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1185–1194. ACM, 2010.

[2] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis.

[3] C. Lu, W. Lam, and Y. Zhang. Twitter user modeling and tweets recommendation based on wikipedia concept graph. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.