The Art of Forgetting Content in Collaborative and Temporal Collections: the case of Wikipedia and Stack Overflow

A lot of research has been invested in the recent years to study how temporal collections evolve over time and how their content is created and organized. However, far less effort has been spent in investigating how the information contained in such collections tends to be forgotten, either explicitly due to deletions or implicitly due to the absence of views. For this study, we are especially interested in two collaboratively edited temporal collections: Wikipedia (along with its revision history) and Stack Overflow. The former can be regarded as a collective memory, where the occurrence of real-world events triggers the creation, editing, aggregation of content by the editors. The latter is a question and answer platform where programmers exchange their knowledge.

Within this context, we want to investigate the following research questions:

- Is any content in temporal collaborative collections forgotten over time?
- What are the contents most subject to forgetting?
- Does the forgetting happen explicitly (via deletions and aggregations) or implicitly (still present but not accessed and viewed)?
- Is the content deletion and aggregation triggered periodically or only due to the arrival of new information (e.g. due to the occurrence of new real-world events)?
- Can we predict which content will be forgotten (either deleted or not viewed) and when?

Requirements

- Good experience with programming languages (such as Java, Python, Matlab, R)
- Knowledge of Machine Learning
- Experience with Machine Learning tools/libraries is a plus

Contacts

If you are interested or have questions, please contact: Andrea Ceroni (email: ceroni@l3s.de).