Method for Novelty Recommendation Using Topic Modelling

Matúš TOMLEIN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
matus@tomlein.org

There is a large number of news and other articles being published on the web by various large and small portals every day. However, the content in these articles is often repeated among them and most of them contain little novel information.

When a person reads an article about a specific topic, it is very likely that they might find dozens of similar articles on the Web, giving the same information in a different way. Such articles are not interesting to the reader. However, there might also be numerous related articles that contain significant novel and interesting information. The problem we deal with is to identify articles with novel and relevant information.

There were several attempts to create news recommender systems that applied novelty detection methods to provide an interface for users to find articles with novel information. They applied various difference metrics for novelty detection, like inverse cosine similarity, Kullback-Leibler divergence, density of previously unseen named entitites, quantifiers and quotes. Using similarity measures as the basis for novelty detection is also common in other works, which mostly focus on sentence-level novelty detection. Sentence-level novelty detection was the main topic of several TREC Novelty track workshops.

So far, the use of topic modelling in novelty detection has not been widely explored. However, there has been a research comparing topic modelling with cosine similarity in novelty detection and it showed promising results in favour of topic modelling [2].

Our method uses topic modelling in order to calculate the novelty and relevancy of articles. Topics are sets of relevant words with some probabilistic degree of distribution with them [2]. We use the Latent Dirichlet Allocation algorithm for topic modelling.

The reason why we think topic modelling can be useful in novelty recommendation is that it provides a way to work with the information in articles on a higher level of abstraction. It allows us to work with information using topics as

* Supervisor: Jozef Tvarožek, Institute of Informatics and Software Engineering

Spring 2014 PeWe Workshop, March 21, 2014, pp. 15–16.

opposed to using keywords. This is particularly useful if we want to track similar information accross articles and find novel groups of information in them.

It is important to ensure that the recommended articles are relevant to the interests of the readers, i.e. to what they previously read about. To achieve this, we cluster articles into groups based on their similarity and recommend novel articles from within the clusters that the user previously read about. We designed a simple method for clustering articles based on their topics. We decided to design and implement our own method because we wanted to make use of our topic model and for its simplicity.

Topics retrieved from LDA have various qualities. Some contain important information, some are just groups of words without significant importance or meaning. These less important topics can have an impact on the performance of our method and so it is useful to give them a lesser importance when considering their contribution.

We also want to give a lesser importance to topics that group information the user already read about. This is a crucial part of our method that ensures the novelty in our recommendation. To meet this goal, we employ topic ranking. We give each topic a numeric rank that represents its importance and novelty to the user. The rank of a topic is calculated separately for each user based on their user model.

We use an algorithm inspired by the method proposed in [1] that calculates the novelty of an article based on the Inverse Document Frequency (IDF) of its terms. We use the average IDF of the 100 best terms of a topic to calculate its rank. As the corpus of documents for calculating the IDF against, we use the articles the user read.

We evaluated our method in a preliminary experiment. The goal of the experiment was to find out the advantages and disadvantages of our method compared to two other commonly used methods for novelty detection.

The experiment went on for a day and a half and 5 subjects (university students) took part in it. They compared 152 pairs of articles. The articles being compared were retrieved from several well-known tech blogs.

We compared our method with two baseline methods used for novelty recommendation: inverse cosine similarity and IDF scored novelty detection [1].

The results were optimistic, giving better results in terms of relevancy of articles and also in recommending articles the participants chose to read next. The method for novelty detection using IDF scoring of terms gave better results in terms of novelty, however their relevancy was poor.

Amended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 185-190.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Karkali, M., Rousseau, F., Ntoulas, A.: Efficient Online Novelty Detection in News Streams.

[2] Sendhilkumar, S., Nandhini, N., Mahalakshmi, G.: Novelty Detection via Toppic Modeling in Research Articles. airccj.org, 2013, pp. 401–410.