UDC 004.912

RECOMMENDER SYSTEM DEVELOPMENT BASED ON TOPIC PROFILES OF OBJECTS

Y. V. Solomonova
(ITMO University, St. Petersburg)
Research Supervisor – M. V. Khlopotov, Ph.D.
(ITMO University, St. Petersburg)

Introduction. Topic profile is a vector of ratings corresponding to some topics [1]. We suggest an approach for recommender system development based on topic profiles of objects that includes topic profile calculation both for the user (based on information from the linked social network account) and for the post on the website. Having calculated these vectors, we can bring the user and the post into a single space, characterized by the values of topic vectors. Thus, recommendations for the user are formed based on comparison of his own topic vector and topic vectors of posts.

The purpose of this work is to develop a recommender system that forms recommendations based on topic profiles of objects.

Interim study results:

In the course of work, we performed a review of seven universal classification systems selected based on GOST 7.59-2003 [2] in order to identify the systems that can be used as a topic profile basis. As a part of the review, we considered the breadth of coverage of different areas of knowledge by classification systems, as well as their redundancy. Based on review results we selected SRSTI as a topic profile basis for further work.

We selected keywords to describe topic profile categories. In order to conduct an automated search for keywords it is necessary to collect a sufficient number of texts uniquely belonging to each of the classifier categories. We chose an electronic catalog of the State Public Scientific-Technical Library of the Russian Academy of Sciences Siberian Branch (SB RAS) as a source for data collection [3]. Due to the fact that the books in the library were divided into the SRSTI classifier categories by experts, we can confidently use collected data as source for keyword selection.

For every SRSTI category, we collected data on book titles, annotations and subject headings using a Python parser. We normalized collected text and automatically selected 200 most popular unigrams and bigrams for each category. We manually checked list items for compliance with category topic and removed ill-suited words from the lists. As a result, we formed a dictionary that stores keywords for each of the 69 SRSTI categories.

We developed an algorithm that calculates text topic profile based on the mentioned keywords frequency. We also developed a function that compares topic vectors based on cosine distance between them.

We also suggested approaches to vector space improvement. We considered the possibility of reducing the vector space by the similar categories merging, as well as improving keywords weight calculation by dividing the keywords weight by the number of its occurrence in dictionary. All suggested approaches were evaluated on the additional catalogs of the library [3]. Improved weight calculation algorithm showed the best results. On the other hand, uniting similar categories only worsened the results.

Main results and future work:

The developed algorithm with improved weight calculation was implemented on the OpinionLine website [4]. The post topic profile is automatically recalculated with each update of its content (quotes and polls related to the post). The user topic profile is calculated based on text data

collected from the linked Vkontakte account (in particular, user groups and subscriptions). Thus, individual post recommendations are formed based on comparison of user and post topic profiles.

As of future work, we plan to conduct an experiment in which the user will be randomly assigned one of two recommendations strategies:

– content-based strategy (recommendations are formed based on user views of posts; the posts classifier contains 10 categories and was developed independently);

– strategy based on the comparison of topic vectors (the SRSTI is the topic profile basis).

Based on the collected data and the various metrics application (such as user retention, diversity and novelty of recommendations, etc.), we will evaluate the compared approaches and draw conclusions on how these approaches impact the user's behavior on the website.

References:

1. Lexin V.A.: Personalization technology based on the identification of topic user profiles and Internet resources, http://www.machinelearning.ru/wiki/images/c/c2/Lexin07master.pdf. Last accessed 25 Feb 2019.

2. GOST 7.59-2003 SSILP. Indexing of Documents. General Requirements for Classifying and Subject Indexing, http://docs.cntd.ru/document/1200032034. Last accessed 26 Feb 2019.

3. State Public Scientific-Technical Library of the Russian Academy of Sciences Siberian Branch (SB RAS), http://webirbis.spsl.nsc.ru. Last accessed 26 Feb 2019.

4. OpinionLine, https://www.opinionline.ru. Last accessed 27 Feb 2019.