The Elicitation of New Users' Interests on Wikipedia

Extended Abstract

Ramtin Yazdanian
EPFL
Lausanne, Switzerland
ramtin.yazdanian@epfl.ch

Leila Zia
Wikimedia Foundation
leila@wikimedia.org

Robert West
EPFL
Lausanne, Switzerland
robert.west@epfl.ch

ABSTRACT

Within the context of recommender systems, the user cold-start problem is the well-known problem of recommending items to a new user with no prior activity in the system. In this paper, we solve this problem in the context of a recommender system for Wikipedia editors. Our method is based on joint topic extraction from the documents' content and the users' editing history, using matrix factorisation and embedding the users and documents in a latent topic space. Questions are extracted from topics - dimensions of the latent space - by taking the 20 highest weighted and 20 highest negatively weighted documents in each topic, and posing the question as "Which set of documents would you be more interested in editing?", with a "neither" answer also being possible. Our results show that our method is capable of creating semantically cohesive questions, and exhibits good performance in offline user-based evaluation.

CCS CONCEPTS

•Computing methodologies → Learning from implicit feedback; Factorization methods;

KEYWORDS

Recommender systems, Cold-start problem, Questionnaire, Wikipedia

1 INTRODUCTION

Recommender systems have become a widespread phenomenon in systems with many users, with one of the well-known examples being Netflix [1]. The aim of a recommender system is to recommend items to users based on the system's understanding of the user's interests, other users' interests, and in some cases the content of the items themselves. The approaches based only on user-user similarities and shared interests are collectively called collaborative filtering (CF) [2], while the approaches only based on the items' content are called content-based filtering [3], with several methods using a hybrid of the two.

Wikipedia is also a system for which recommender systems can be developed in two ways: for readers, and for editors. We only consider the latter case. In particular, we discuss the user cold-start problem [4], in which recommendations are to be made for a new user, for whom no past history exists. This poses a challenge to recommender systems, because they rely on the user's past history to recommend related items. As we will elaborate further in the related work section, many of the methods used to solve this problem are based on questionnaires posed to the user, by means of which the system creates an initial profile for the user and then recommends items based on this profile. The main problem to be solved then, is how the questionnaire should be generated from the data. Creating a recommendation system for Wikipedia poses two challenges in particular:

(1) Wikipedia editors do not "rate" articles; the only information available about their preferences is their editing history. This is known in the literature as implicit feedback [5].
(2) As opposed to most systems which recommend "consumable" items such as movies to watch, recommending Wikipedia articles to editors would not be for their consumption, but for their contribution; and while every movie can be watched, not every Wikipedia article needs contributions (and additionally, not every user will be qualified to contribute to a given article).

In this article, we propose a method to create a questionnaire for the user (i.e. editor) cold-start problem in Wikipedia, dealing with both of the aforementioned problems. Our contribution is twofold: we propose a method for extracting topics jointly from the users' editing history and the articles' content by embedding both in a latent space, and we propose a question creation scheme by converting each of the extracted topics into comparative questions of the form "Would you be more interested in article set A or article set B?", with users choosing either one of the two sets, or neither.

2 OUR WORK

2.1 The data

Our data includes implicit feedback in the form of edit counts, which we calculate using the revision histories of Wikipedia articles, and also the content of the said articles.

2.2 Method summary

We create a document-term matrix from the documents' content, using a bag of words representation of documents which is then further refined through TF-IDF weighting. In addition, we create a user-document matrix whose entries are the number of edits made by a user to a document. Our method uses singular value decomposition (SVD) to factorise the document-term matrix first, enabling us to embed the documents in a k-dimensional latent space by taking the linear subspace spanned only by the singular vectors corresponding to the k highest singular values. We then proceed to use this embedding as a prior for a joint embedding of both documents and users into a latent space by factorising the user-document matrix, while striving to keep the latent document representation close to the result of our SVD. According to our observations, topic extraction only based on the documents' contents

will be noisy, whereas only using the (in our experience) more noisy editing history will result in topics that seem irrelevant due to the editing matrix also capturing information such as the controversial nature of articles or simply their length. However, used together, the resulting topics will be relatively similar to the content-based approach, but with greater cohesion, which we quantify.

Once documents have been embedded in the latent space, for each of the dimensions of this space we proceed to take the 20 highest positively weighted (top 20) and 20 highest negatively weighted documents (bottom 20). Our idea is that these documents will more or less embody the main word co-occurrence and editor co-occurrence patterns captured by that dimension of the latent space, and in opposing ways; therefore by posing each question as "Would you be more interested in editing set A or set B?", a user's answers to them would define an initial profile for said user. This allows us to pair the user with another newcomer based on their answers, thus empowering and encouraging them to contribute.

2.3 Results

Our results so far indicate that the latent document representation matrix (Q, whose rows are documents) produced by our full method does indeed produce better questions than an SVD on the document-term matrix ($\bar{Q}$, again rows being documents), as demonstrated in fig. 1, which shows a comparison of cohesion scores for the first 50 questions generated from Q and a reduced version using $\bar{Q}$. To calculate the cohesion scores, the questions are produced in one case by our full method and in the other by an SVD, and then for each question, the average of cosines among the $\bar{Q}$ representations of the top 20 documents is summed up with the average of the cosines among the bottom 20 documents. We have used the scheme to facilitate comparison.

We also perform an offline experiment by simulating the questionnaire-answering process by a set of test users after having hidden 20 of the documents they have edited. Having their simulated initial profiles, we create a graph in which each node is one of the test users, and we solve a maximum-weight max-cardinality matching problem using the Hungarian algorithm (using the number of questions they have the same answer to as the weight for the edge connecting their respective nodes in the graph). Afterwards, within each pair, we take the latent representations (according to Q) of the 20 hidden documents of each of the users in the pair, and calculate the average pairwise cosine of all pairs of one document from each user (we call this scheme "soft scoring"). Fig. 2 shows the distributions of these scores for pairs created using the aforementioned method, versus random pairwise matching.

2.4 Future work

Our online experiments will be based on pairing real users together, and will involve users taking the questionnaire, then being paired together, and finally giving us feedback on what they thought of their paired partners after having had a chat with them. This will be part of a greater experiment alongside Wikimedia, in the framework of an integrated recommender system and aimed at improving work in a voluntary environment.

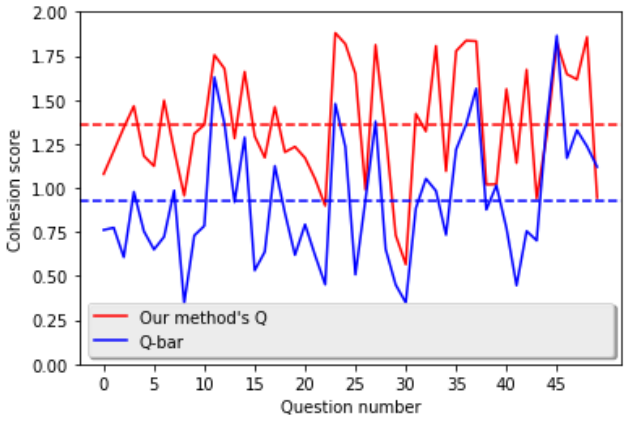

Figure 1: Comparison of cohesions for for the first 50 questions generated from Q and $\bar{Q}$. Average cohesion for each is shown as a dashed horizontal line.

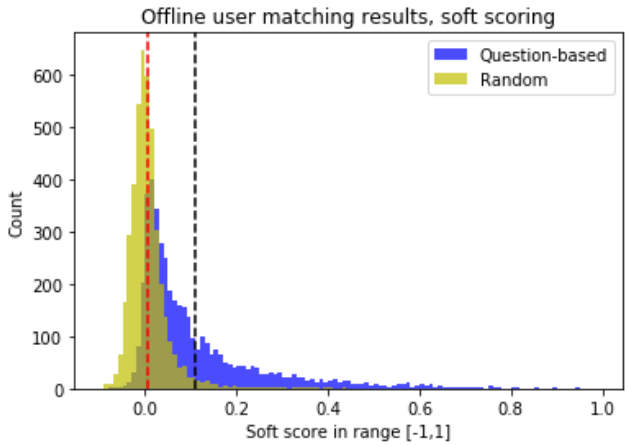

Figure 2: Comparison of offline performance for matching based on simulated answers to our questions, versus random matching. Vertical lines are averages over all pairs.

REFERENCES

[1] J. Bennett and S. Lanning, "The Netflix Prize". Proceedings of KDD Cup and Workshop 2007, San Jose, California, Aug 12, 2007.
[2] D. Goldberg, D. Nichols, B. M. Oki and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry". Communications of the ACM 35 (1992), 61fi?!70.
[3] R. J. Mooney and L. Roy. "Content-based book recommending using learning for text categorization". Proceedings of the Fifth ACM Conference on Digital Libraries, pages 195-204, 2000.
[4] A. I. Schein, A. Popescul, L. H. Ungar, D. M. Pennock, "Methods and Metrics for Cold-Start Recommendations". Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, August 2002, p. 253-260.
[5] Y. Hu, Y. Koren, C. Volinsky, "Collaborative Filtering for Implicit Feedback Datasets". Eighth IEEE International Conference on Data Mining (ICDM), 2008.