The FairGRecs Dataset: A Dataset for Producing Health-related Recommendations

Maria Stratigi[1], Haridimos Kondylakis[2], and Kostas Stefanidis[3]

[1] University of Tampere, Tampere, Finland, Maria.Stratigi@uta.fi
[2] ICS-FORTH, Heraklion, Greece, kondylak@ics.forth.gr
[3] University of Tampere, Tampere, Finland, kostas.stefanidis@uta.fi

Abstract. Nowadays, the number of people who search for information related to health online has significantly increased, while the time of health professionals for recommending useful sources online has been reduced to a great extend. As such, providing valuable information to users for health related issues, based on their personal health profiles, in the form of suggestions, approved by their caregivers, can significantly improve the opportunities that users have to inform themselves online about health problems and possible treatments. However, due to several legal and ethical constraints, personal health profiles usually are not accessible. In this paper, we present FairGRecs, a synthetic dataset that can be used for evaluating and benchmarking methods that produce recommendations related to health documents based on individual health records. Specifically, FairGRecs can create, via a fully parametrized API, synthetic patients profiles, containing the same characteristics that exist in a real medical database, including both information about health problems and also relevant documents.

1 Introduction

Medicine is undergoing a revolution that is transforming the nature of health-care from reactive to preventive. One of these challenges in this revolution is the problem of the quality and the amount of information that can be found online [1], since health information is one of the most frequently searched topics on the Web. Especially during the last decade, the number of users who look online for health and medical information has dramatically increased. However despite the increase in those numbers and the vast amount of information currently available online, it is very hard for a patient to accurately judge the relevance of some information to his/her own case and to identify the quality of the provided information.

Furthermore, the optimal solution for patients is to be guided by healthcare providers to appropriate sources of information [1], [13]. Delivering accurate information sources to a patient, increases his/her knowledge and changes the way of thinking, which is usually referred as patient empowerment [7], [8]. As a result, the patient's dependency for information from the doctor is reduced. Also, patients feel autonomous and more confident about the management of their disease [16]. To this direction, health providers have the history of their patient's and their interests, in order to make an informed decision about the information that would likely be beneficial for them. However, health

providers have less and less time to devote to their patients. As such, guiding each individual patient appropriately is a really difficult task.

On the other hand, the use of group-dynamics-based principles of behavior change have been shown to be highly effective in enhancing social support through promoting group cohesion in physical activity [2], in reducing smoking relapse [3] and in promoting healthy dietary habbits [10]. In small groups, therapy sessions enjoy a social component as participants can share experiences and discussion. In those therapy sessions, a caregiver can guide patients to more optimal resources over the Web. However, if identifying online information content for a single patient is a difficult task, identifying information for a group of participants is a really challenging one.

To this direction, in our work [14, 15], we focused on recommending interesting health documents, to groups of users. Our motivation was to offer a list of recommendations to a caregiver who is responsible for a group of patients. The recommended documents need to be relevant, based on the patients current profile, namely by exploiting the patients personal healthcare record (PHR) data. However, although it is really common for patients to look for health information and sometimes to rate related documents on the Web, their profiles are usually not accessible, neither linked to those documents. Among others, legal and ethical constraints prohibit the collection and the exploitation of such a dataset.

This way, in this paper, we present a synthetic dataset, FairGRecs, that can be used for evaluating and benchmarking methods that produce recommendations related to health documents. More specifically, we rely on the EMRBots dataset[4], which contains synthetic patients profiles, containing the same characteristics that exist in a real medical database, such as patients admission details, demographics, socioeconomic details, labs and medications, extending it with a document corpus and a rating dataset. By exploiting the FairGRecs dataset, interested users can create patients that have provided rankings for health documents. To link document contents with patients, we use the ICD10[5] ontology, namely the International Statistical Classification of Diseases and Related Health Problems, which is a standard medical classification list maintained by the World Health Organization. FairGRecs is fully parametrized and is offered via an API.[6]

This dataset has been used already for optimizing the Personal Health Information Recommender (PHIR) [5], [8], [9], developed within the EU project iManageCancer [6]. PHIR is a recommendation engine for recommending high quality cancer documents selected by health providers to patients.

The rest of this paper is structured as follows. Section 2 introduces our synthetic dataset, consisting of patients data, a document corpus and a ratings dataset. We also include a dataset creation example to showcase how to build these datasets. Section 3 describes the developed application programming interface, while Section 4 concludes with a discussion on the usefulness of the produced dataset.

[4] http://www.emrbots.org

[5] http://www.icd10data.com/

[6] https://bitbucket.org/MariaStratigi/fairgrecs-dataset/overview

2 The FairGRecs Dataset

In general, a recommender system requires access to a certain amount of information – set of medicinal documents, a dataset containing ratings that the users have given to those documents and the personal health information of the users, to make suggestions to users. The absence of real data, motivated us to develop a fully parameterized tool that automatically creates the necessary datasets, that we do not have access to.

The first major obstacle is the acquisition of the dataset containing the personal health information of the users. Among others, legal and ethical constraints prohibit the collection and the exploitation of such a dataset. A further constraint is that the health information dataset has to be linked to a document corpus via a users' ratings dataset. So the problem we face is the need of three interlinked datasets, that are void of any legal and ethical constrains. The first step to overcome this obstacle, is the adoption of the 10.000 chimeric patient profiles provided by EMRBots.

2.1 Patient Profiles Dataset

Unlike other methods that typically obscure or shift real patients' data elements, the EMRBots proposed methodology is invulnerable in terms of security, because it does not rely on real data elements pulled from an existing *Electronic Medical Record* (EMR); therefore, it is not associated with privacy concerns in regard to individuals' sensitive data.

The data is generated according to pre-defined criteria and is not based on any human data. These criteria are divided into *Population-level* and *Patient-level* characteristics. The first group offers an array of values in order to define demographics, such as gender, marital status, major language, ethnicity, date of birth and income level. After the completion of the *Population-level* configuration, there are n patients generated. For each such patient, the *Patient-level* configuration associates with them additional details. These details include randomly generated length of stay (in days) and start and end dates. Furthermore, each admission is associated with laboratory measurements and a chief complaint randomly selected from a list of International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes. In the last step of the patient-level configuration, laboratory values are added. Laboratory values are based on 35 common types - for instance, sodium levels, creatinine levels, or platelet count.

2.2 Document Corpus & User Preferences Dataset

One of the most important information a recommender system requires is a dataset containing the users preferences towards a set of items [12]. These preferences can take many formats, such as ratings, check-ins or textual reviews [11]. For this work, our chosen format is that of a ratings dataset. Most specifically, we produce a dataset containing ratings associated with users for particular health related documents.

Based on the patients' health profiles dataset that we have already acquired, we generate two new datasets. The first is a document corpus that consist of documents' id codes and their corresponding keywords, and the second is the ratings dataset, that incorporates the ratings given by the patients to the documents.

The generation process for both of these datasets, is fully parameterized. It is worth mentioning that, because our main focus is the development of the ratings dataset, heavy emphasis is given to the health profile of the patients. This is mostly presented as the health problems of each patient. As mentioned in the previous section this information is documented using ICD-10 codes.

Document Corpus. The generation of the document corpus, includes a document id and the corresponding keywords for each created document. As such we are not generating full text documents. To achieve that, we take advantage of the ICD-10 ontology tree. We generated a $numDocs$ number of documents, for each second level category of the ICD-10 ontology (i.e., for each node that belongs in the second level of the ontology tree). We will call these *primary nodes*. The ICD-10 ontology tree has 295 such nodes. Here, we make the assumption that a document cannot be related to more than one category. For example, a document cannot impart information for pregnancy and the perinatal period, cause these areas are represented by two different *primary nodes* in the tree.

For the documents' corresponding keywords, we randomly selected $numKeyWords$ words from the description text of the nodes in each subsequent subtree of the *primary node* that the document belongs. From those descriptions, we removed the most common words – often called stopwords, such as "the", "and","it" and made our selection from the rest. A visual description is provided in Figure 1.

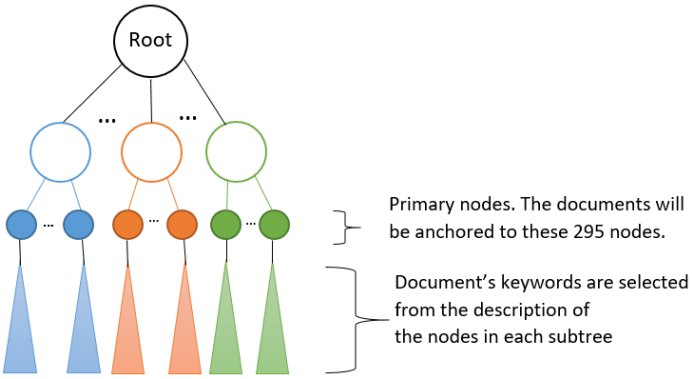

Fig. 1. The created documents will be anchored to the primary nodes, while their keywords will be selected from the subtree.

When ranking items based on human preferences, the most common observed distribution of these ratings is the power law distribution [4]. In order to make our dataset as plausible as possible, we depict this distribution by randomly selecting a $popularDocs$ number of documents that will be the most popular. These documents span through all the categories and are selected arbitrarily.

Ratings dataset. In order to create the ratings dataset, first we assume that all the patients have given at least a minimum non zero number of ratings. This assumption was made in order to make our datasets easier to work within the domain of recommender systems, and more specifically the systems that operate under a collaborative filtering design. In general these systems operate by finding similar users to a given user, and extrapolating an item's rating based on the scores given by those users [14, 15]. If a user has not given any ratings, we will stumble upon the cold start problem, where we cannot find any similar users to him/her and subsequently, we will not be able to provide him/her with any recommendations.

In order to avoid this problem, we have surmised that all patients have given a $numRatings$ number of ratings. Specifically, we have divided the patients into three groups – $occasional$, $regular$ and $dedicated$. The users in each group have given *low*, *average* and *high* number of ratings, respectively. The number of ratings are randomly selected from a numerical range, based on which of the three groups the patient belongs.

Based on that number, we will create a corresponding number of rating nodes for each user. A rating node links a user to a document, i.e., the user has rated that document. A user cannot have more than one rating node for the same document, but can have many nodes for different documents that belong in the same category.

We have divided the user's rating nodes into two groups; $healthRelevant$ and $nonRelevant$. Using the health problems data (noted in ICD-10 nodes) of each user, the first group of ratings will go to documents belonging to the same subtree as one of their health problems, while the second group will be randomly assigned to the rest of the documents. Our assumption here is that the patients will be interested not only in documents regarding their health problems, but also to some extent in others as well.

Finally, in the last step, we assign rating values for each rating node that we generated previously. We choose the nodes randomly, and assign to them *a value* in the range of 1 to 5. The user is able to define the total number of ratings with a specific value that will be present in the rating dataset. This is accomplished as shown before with the use of percentages. For example, the administrator can define that the 25% of all ratings will have the value of 1.

2.3 Datasets Creation Example

In Section 2.2, we analyzed the proposed method of creating two datasets - documents and ratings - in lieu of real data. In Tables 1 and 2, we present the parameters needed for creating an example dataset and we briefly explain the values given to them. Furthermore, we selected to use the 10.000 patients chimeric dataset provided by the EMR-Bots. After all the necessary steps were completed the number of items in the document corpus was 79.650 and the total number of ratings generated was 1.576.872.

Figure 2 depicts the distribution of ratings in the documents. We partition the ratings in groups of 50. Most of the documents (71%) have received ratings in the range of [50-100]. In the second place (21%), we have the documents that have been rated from 0 to 50 times, while if we accumulate all the documents which have been rated more than 200 times, they merely make up of the 1.12% of the corpus. As expected, these results simulate a power law, where the prominent items are few, and the plethora of documents have very low popularity.

Table 1. The parameters needed to creating the document corpus.

Parameter Name	Explanation	Value
numDocs	The number of documents created for each different category of health problems, based on the ICD10 ontology tree.	270
numKeyWords	The number of randomly selected keywords, attached to each document.	10
popularDocs	The number of documents, that will be most popular in each category, in order to simulate a power law distribution.	70

Table 2. The parameters needed to create the ratings dataset.

Partitions	Parameter Name	Explanation	Value
Groups	Group $occasional$	Users give [20,100] ratings	50% of all patients
	Group $regular$	Users give [100,250] ratings	30% of all patients
	Group $dedicated$	Users give [250,500] ratings	20% of all patients
Scores	One	Ratings valued as 1	20% of all ratings
	Two	Ratings valued as 2	10% of all ratings
	Three	Ratings valued as 3	30% of all ratings
	Four	Ratings valued as 4	20% of all ratings
	Five	Ratings valued as 5	20% of all ratings
Ratings	healthRelevant	Ratings relevant to health problems	20% of user's ratings
	nonRelevant	Ratings not relevant to health problems.	80% of user's ratings

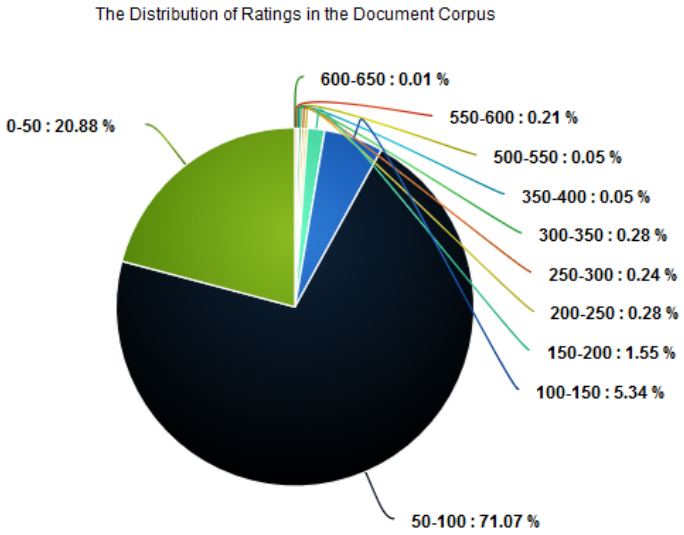

Fig. 2. The distribution of the ratings in the document corpus, where we partition the number of ratings in groups of 500.

3 Application & Programming Interface

In order to streamline the creation process of the datasets and make it more user friendly, we developed an application and the corresponding Application Programming Interface (API). Those incorporates all the functions described in the previous section. They were both developed using Java, and they are available online [7]. The application interface, shown in Figure 3, is divided into four main tabs, each dedicated to a different part of the datasets creation process.

The first tab called 'File Paths', is a form for loading the two EMRbots dataset files, that contain the patients basic information and their health problems. In addition, the ICD-10 ontology is needed in a xml format, as well as a stopwords file containing the most common words in the English language.

The second tab is about the document corpus. The user needs to enter a numerical value, regarding the number of documents that will be created per category (i.e., *first level nodes*), and the number of keywords each document will have. The final input concerns the number of documents that will be popular per category, in order to simulate a power law distribution.

In the third tab, we have accumulated all the parameterized variables regarding the patients. These predominantly concern the patients partitioning into groups. Specifically, the user can divide the patients into the three groups by setting the percentage of the patients that belong in each group. The user is also able to define the minimum and maximum number of ratings per different group. Finally, the distribution of the user's

[7] `https://bitbucket.org/MariaStratigi/fairgrecs-dataset/overview`

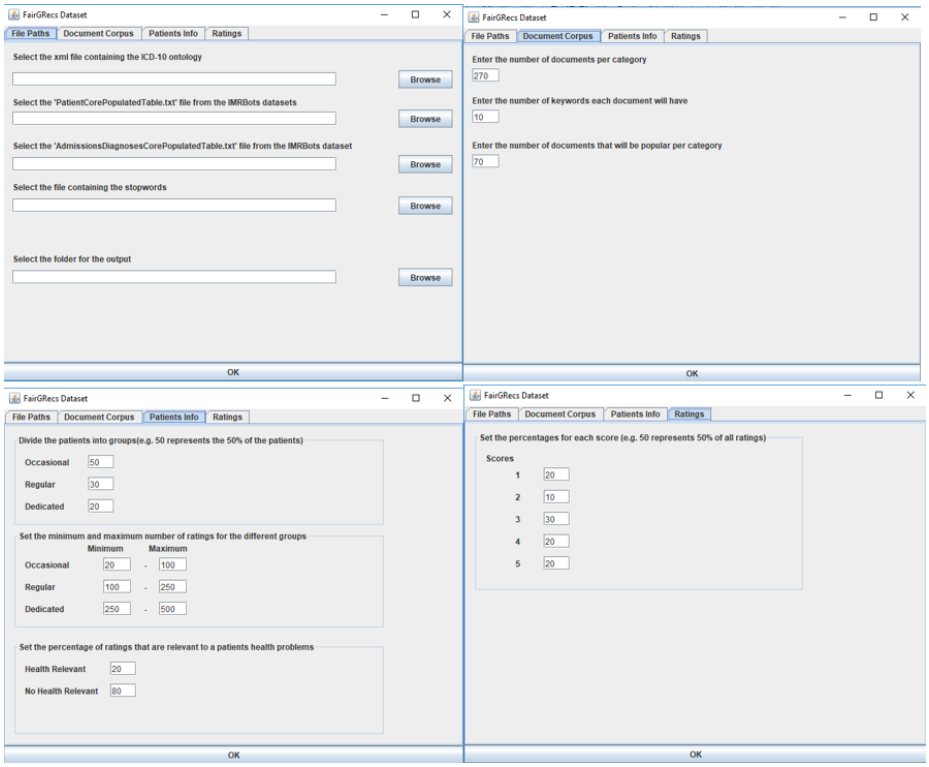

Fig. 3. Visualization of the four different tabs of the API. As input, we have entered the default numbers.

ratings to documents that are relevant to his/her health problems is again accomplished by the use of percentages.

The last tab concerns the distribution of values to the ratings nodes. As before, the user defines the percentage of ratings with a specific value. The selection of a document to assign a value is done randomly.

4 Discussion & Conclusion

To the best of our knowledge, the FairGRecs dataset, is currently the only dataset available, combining personal health record, documents and ratings, offering a unique opportunity for experimenting with recomender systems in the health domain. As such it fills an important gap in the area of health recomender systems and it reuses and significantly extends state of the art datasets. Furthermore, is of particular interest to the Semantic Web Community, as in its core a widely used taxonomy is used, opening many possibilities for subsequent exploitation through reasoning techniques.

In addition, there is significant evidence of usage by the community of health recommender systems, as currently more and more platforms emerge enabling patients

to access high quality health related information. Nevertheless, we do not only offer a specific dataset but also an application and the corresponding API, enabling experimentation with endless possibilities, as well as its wider adoption and extensibility. The source code is also available allowing further extensions by the community.

References

1. G. M. Berg, A. M. Hervey, D. Atterbury, R. Cook, M. Mosley, R. Grundmeyer, and D. Acuna. Evaluating the quality of online information about concussions. *Journal of the American Academy of PAs*, 27:1547–1896, 2014.
2. I. Brandon, K. Daniel, C. Patrice, and T. Nicholas. Testing the efficacy of ourspace, a brief, group dynamics-based physical activity intervention: A randomized controlled trial. *J Med Internet Res*, 18(4):e87, May 2016.
3. D. Y. T. Cheung, H. C. H. Chan, J. C.-K. Lai, V. W. F. Chan, P. M. Wang, W. H. C. Li, C. S. S. Chan, and T.-H. Lam. Using whatsapp and facebook online social groups for smoking relapse prevention for recent quitters: A pilot pragmatic cluster randomized controlled trial. *J Med Internet Res*, 17(10):e238, Oct 2015.
4. H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 211–220, New York, NY, USA, 2007. ACM.
5. G. Iatraki, H. Kondylakis, L. Koumakis, M. Chatzimina, K. Marias, and M. Tsiknakis. Personal health information recommender: implementing a tool for the empowerment of cancer patients. *ecancer 12 851*, 2018.
6. H. Kondylakis, A. I. D. Bucur, F. Dong, C. Renzi, A. Manfrinati, N. M. Graf, S. Hoffman, L. Koumakis, G. Pravettoni, K. Marias, M. Tsiknakis, and S. Kiefer. imanagecancer: Developing a platform for empowering patients and strengthening self-management in cancer diseases. In *IEEE CBMS*, pages 755–760, 2017.
7. H. Kondylakis, L. Koumakis, E. Genitsaridi, M. Tsiknakis, K. Marias, G. Pravettoni, A. Gorini, and K. Mazzocco. Iems: A collaborative environment for patient empowerment. In *12th IEEE International Conference on Bioinformatics & Bioengineering, BIBE*, pages 535–540, 2012.
8. H. Kondylakis, L. Koumakis, E. Kazantzaki, M. Chatzimina, M. Psaraki, K. Marias, and M. Tsiknakis. Patient empowerment through personal medical recommendations. In *MEDINFO*, page 1117, 2015.
9. H. Kondylakis, L. Koumakis, M. Psaraki, G. Troullinou, M. Chatzimina, E. Kazantzaki, K. Marias, and M. Tsiknakis. Semantically-enabled personal medical information recommender. In *ISWC*, 2015.
10. J. Meng, W. Peng, Y. S. Shin, and M. Chung. Online self-tracking groups to increase fruit and vegetable intake: A small-scale study on mechanisms of group effect on behavior change. *J Med Internet Res*, 19(3):e63, Mar 2017.
11. E. Ntoutsi and K. Stefanidis. Recommendations beyond the ratings matrix. In *Proceedings of the Workshop on Data-Driven Innovation on the Web, DDI@WebSci 2016, Hannover, Germany, May 22-25, 2016*, pages 2:1–2:5, 2016.
12. E. Ntoutsi, K. Stefanidis, K. Rausch, and H. Kriegel. Strength lies in differences: Diversifying friends for recommendations through subspace clustering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 729–738, 2014.
13. S. Schaller, V. Marinova-Schmidt, M. Setzer, H. Kondylakis, L. Griebel, M. Sedlmayr, E. Graessel, M. J. Maler, S. Kirn, and L. P. Kolominsky-Rabas. Usefulness of a tailored

ehealth service for informal caregivers and professionals in the dementia treatment and care setting: The ehealthmonitor dementia portal. *JMIR Res Protoc*, 5(2):e47, Apr 2016.
14. M. Stratigi, H. Kondylakis, and K. Stefanidis. Fairness in group recommendations in the health domain. In *ICDE*, 2017.
15. M. Stratigi, H. Kondylakis, and K. Stefanidis. FairGRecs: Fair group recommendations by exploiting personal health information. In *DEXA*, 2018.
16. M. Wiesner and D. Pfeifer. Adapting recommender systems to the requirements of personal health record systems, 2010.