

І.І. Оберемок; Київ. нац. ун-т буд-ва та архіт. МОН України. – К., 2003. –17 с. 6. *Оборский Г.О.* Стандартизация і сертифікація процесів управління якістю освіти у вищому навчальному закладі / *Г.О. Оборський, В.Д. Гогунський, О.С. Савельєва*/ Праці Одеського політехнічного університету, 2011, Вип. 1(35). – С. 252–256. 7. *Онищенко С.П.* Основные объекты маркетинга в проектной деятельности // *С.П.Онищенко, Т.И.Берневек* // Восточно-Европейский журнал передовых технологий, 2013. – №3/2(63). – С.8-12. 8. *Логинов О.В.* Современные подходы к созданию проектно-ориентированного вуза / *О.В.Логинов* // Восточно-Европейский журнал передовых технологий, 2013. – №1. – С.137–139. 9. *Онищенко С.П.* Моделирование процессов организации и функционирования системы маркетинга морских транспортных предприятий / *Онищенко С.П.* – Одесса : «Феникс», 2009. – 328 с

**Bibliography (transliterated):** 1. Dolyatovskiy, et al. V.A. *Strategicheskoye planirovaniye deyatel'nosti vuza na rynke obrazovatel'nykh uslug. Monografiya.* Pod red. V.A. Dolyatovskogo, O.A.Mazura. Rostov-na-Donu-Nevinnomyssk: SKNTS VSH-RGEU «RINKH»-NIEUP, 2005. Print. 2. Tkachuk, S.V. "Profilyuvannya tsinnosti proektiv osvity dlya navchal'nykh zakladiv". *Shlyakhy realizatsiyi kredytno-modul'noyi systemy orhanizatsiyi navchal'noho protsesu i testovyykh form kontrolyu znan' studentiv: Materialy nauk.-metod. Seminaru.* Pid red. Hohuns'koho V.D. Vol. 4: *Orhanizatsiya navchal'noho protsesu i testuvannya.* – Odessa : Nauka i tekhnika, 2010. 58–63. Print. 3. Toplyna, Y.Y. "Povyshenye konkurentosposobnosti vuza y eho razvytye v pozytsyyi marketynha". *Vek kachestva.* No. 1. 2011. 32–33. Print. 4. Kolyada, O.P. "Proyektno-oriyentirovannaya formalizatsiya strategicheskogo komponenta funktsional'noy deyatel'nosti vysshogo uchebnogo zavedeniya". *Upravlinnya proektamy ta rozvytok vyrobnytstva: zb.nauk.pr.* Lugansk: vyd-vo SNU im. V.Dalya, No. 3(27), 2008. 81– 87. 5. Oberemok, I.I. *Metody ta zasoby proyektno-oriyentovanoho upravlinnya u vyshchykh navchal'nykh zakladakh: Avtoref. dys.kand. tekhn. nauk: 05.13.22/ I.I. Oberemok; Kyiv. nats. un-t bud-va ta arkhitekt. MON Ukrainy.* Kiev, 2003. Print. 6. Oborskiy, H.O., et al. "Standartyzatsiya i sertyfikatsiya protsesiv upravlinnya yakisty osvity u vyshchomu navchal'nomu zakladi". *Pratsi Odes'koho politekhnichnoho universytetu,* 2011, No. 1(35). 252-256. Print. 7. Onishchenko S.P., and T.I.Bernevek "Osnovnyye ob'yekty marketinga v proyektnoy deyatel'nosti " *Vostochno-Yevropeyskiy zhurnal peredovykh tekhnologiy,* No. 3/2(63). 2013. 8–12. Print. 8. Loginov O. V. "Sovremennyye podkhody k sozdaniyu proyektno-oriyentirovannogo vuza". *Vostochno-Yevropeyskiy zhurnal peredovykh tekhnologiy,* No. 1. 2013. 137–139. Print. 9. Onishchenko S. P. *Modelirovaniye protsessov organizatsii i funktsionirovaniya systemy marketinga morskikh transportnykh predpriyatiy.* Odessa: «Feniks», 2009. Print.

*Посмутила (received) 28.11.2014*

УДК 004.9

**К. В. МЕЛЬНИК**, ассистент НТУ «ХПИ», г. Харьков

## **ПРИМЕНЕНИЕ АЛГОРИТМА КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ ДЛЯ ОБРАБОТКИ МЕДИЦИНСКИХ ДАННЫХ**

Рассмотрена проблема повышения эффективности работы медицинского учреждения в рамках реализации социального проекта. Выполнен обзор методов коллаборативной фильтрации. Для расчета меры схожести между медицинскими картами пациентов предлагается использовать коэффициент Гауэра.

© К. В. Мельник, 2015

**Ключевые слова:** обработка медицинских данных, коллаборативная фильтрация, коэффициент Гауэра.

**Введение.** Стратегия развития любого государства предполагает улучшение экономической, научно-технической, военной и социальной ситуации в государстве. В современном обществе по разным причинам не всегда уделяется достаточно внимания актуальным социальным проблемам, несмотря на то, что социальная политика также важна, как и развитие военной инфраструктуры. Только государство может с максимальной эффективностью повлиять на разные социальные проблемы. Социальная политика представляет собой систему мер, направленных, например, на следующие актуальные проблемы: вопросы повышения качества жизни населения, проблемы здравоохранения, предоставление качественных образовательных услуг, вопросы социальной справедливости.

Рассмотрим проблему повышения уровня здоровья населения на определенной территории. На здоровье пациентов действует много факторов: экология, наследственность, образ жизни, личная гигиена, качество предоставляемых медицинских услуг, профилактика заболеваний. Часть факторов являются управляемыми, например, качество медицинского обслуживания или профилактика заболеваний. Улучшение этих двух факторов приведет к повышению эффективности работы медицинского учреждения, что, в свою очередь, повысит уровень здоровья в конкретном регионе. Рассматриваемая задача повышения уровня здоровья относится к классу социальных проектов.

**Постановка задачи.** В общем виде социальный проект представляет собой совокупность действий, результатом которых будет уникальный продукт, услуга или информация, которые улучшают социальную ситуацию в конкретном регионе или решают конкретную социальную проблему [1]. В данном случае результатом социального проекта является обеспечение определенного уровня здоровья населения в условиях ограниченного финансирования. Существуют разные варианты решения поставленной задачи, например, создание более выгодных условий работы для медработников, закупка современного оборудования. Но такие способы требуют дополнительного финансирования, которое обычно отсутствует. С другой стороны, задачу обеспечения уровня здоровья можно рассмотреть с точки зрения оказания профилактических услуг в медицинском учреждении. Одна из мер профилактики – это проведение скрининговых процедур, которые включают в себя оценку риска развития заболеваний. Оценить вероятность или риск возникновения заболевания у пациента можно на основе данных из медицинских карточек пациентов. Одно из скрининговых мероприятий – это проведение диспансеризации, которая подразумевает ежегодное комплексное обследование пациентов. Достаточно большая часть

пациентов, которые обслуживаются в конкретном медицинском учреждении, проходят все необходимые процедуры. А это, в свою очередь, приводит к тому, что в карточках накапливается очень много информации, которую потом можно обработать и сделать различные выводы. Например, можно выявить те или иные факторы риска развития заболеваний, или увидеть степень распространенности какого-либо заболевания. Для того чтобы применять полученные выводы для тех пациентов, которые не посещают регулярно медицинские учреждения, необходимо рассматривать только таких пациентов, которые весьма схожи по своему анамнезу. Отбор представительной выборки пациентов можно произвести с помощью техники коллаборативной фильтрации (КФ) [2, 3].

**Обзор методов коллаборативной фильтрации.** Анализ различных источников информации показал, что на сегодняшний момент методы, используемые для КФ, разделились на три основных группы (см. рис. 1).

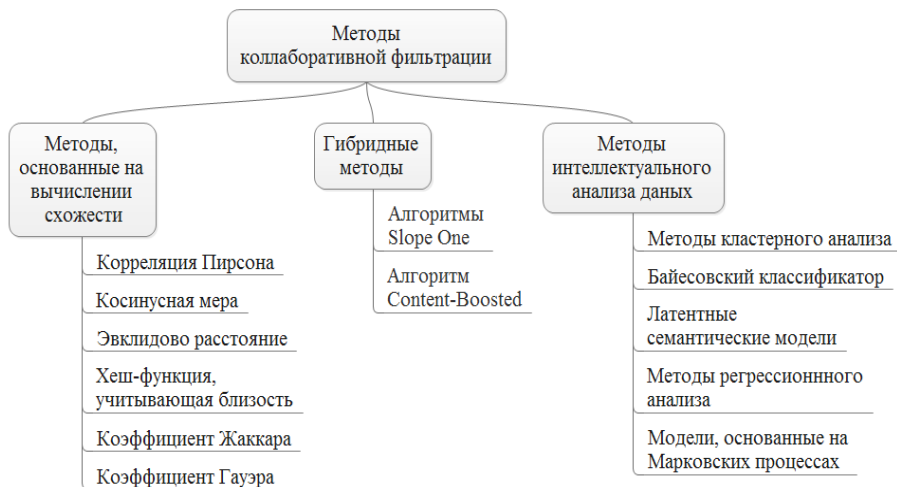


Рис. 1 – Классификация методов коллаборативной фильтрации

Первая группа вычисляет разные меры схожести или подобия между рассматриваемыми объектами. Самыми распространенными считаются вычисление корреляции Пирсона и косинусной меры подобия. Вторая группа методов представляет собой методы интеллектуального анализа данных, что предполагает использование разных методов машинного обучения. Выбор метода в данном случае зависит от предметной области, имеющегося статистического материала и мощности вычислительной системы. Последняя группа методов – это гибридный подход, который подразумевает использование методов, как из первой группы, так и из второй, применяя их в разных комбинациях. Самыми известными считаются группа алгоритмов

Slope One [2], которые основаны на использовании модифицированной формулы регрессионной модели, а также алгоритм Content-Boosted [3], который представляет собой использование байесовского классификатора наравне с применением коэффициента Пирсона.

В работе [4] рассмотрено применение одного из методов КФ, а именно аппарата Байесовых сетей, для обработки медицинских данных. Авторы показали, что Байесовский классификатор – хороший математический аппарат, который можно применять для нахождения оценок рисков заболеваний пациента. Но если в медкарте мало информации, то данный подход характеризуется достаточно большой долей субъективизма.

Рассмотрим алгоритм применения группы методов КФ, основанной на вычислении схожести между медкартами пациентов. На первом этапе происходит оценка данных в медкарте пациента. Если информации в карте достаточно, то сразу вычисляются риски различных заболеваний, в противном случае необходимо применить один из методов КФ. Далее производится выбор способа вычисления меры схожести и последующий расчет выбранной меры. Третий шаг позволяет классифицировать пациентов, а потом выбрать наиболее близких к рассматриваемому. Для этого этапа можно использовать заранее заданный порог схожести или определить предельное число пациентов, меры схожести которых близки к рассматриваемым пациентом. Последний этап – вычисление оценок рисков заболеваний.

Техника КФ дает хорошие результаты на бинарных данных. Обработка медицинских данных сопряжена с рядом трудностей: данные имеют разную природу [5]. Информация в медкарте представлена в качественном и количественном виде. Качественная информация, как правило, представлена либо в бинарном виде, либо порядковым или в шкале категорий. Обычные меры схожести, например, корреляция Пирсона, в случае медицинских данных не применимы. Существуют различные способы обработки данных разной природы. Например, каждый симптом, диагноз или анализ можно представить в бинарном виде: наблюдается или отсутствует; или все данные представить с помощью шкалы категорий, а после преобразования данных применить соответствующие меры схожести. Также можно каждый тип медицинской информации обработать своим способом, а затем вычислить общую меру схожести, используя тот или иной способ свертки. В данной статье предлагается использовать в качестве меры схожести коэффициент Гауэра, который позволяет использовать вместе разные типы данных [6].

**Результаты. Модель оценки рисков развития заболеваний.** Рассмотрим применение метода КФ для вычисления оценок рисков развития заболеваний.

Пусть  $I$  – множество видов записей в медицинской карте, например, факторы риска, диагнозы, симптомы, результаты анализов, результаты

приборно-компьютерных исследований. Тогда  $J_i, i \in I$  – множество записей  $i$ -вида, например, пациент может иметь нарушения сердечно-сосудистой системы и проблемы с печенью, что представляет собой информацию о диагнозах.

Обозначим  $P$  – множество всех пациентов, которые обслуживаются в медицинском учреждении. Тогда  $r_{kij} (k \in P, j \in J_i, i \in I)$  – значение  $j$ -записи  $i$ -вида, представленной в медицинской карте у  $k$ -го пациента ( $k \in P$ ).

Пусть  $w_{ij}^{km}$  – весовой коэффициент важности  $j$ -ой записи  $i$ -го вида двух пациентов  $k$  и  $m$  ( $k, m \in P$ ). Коэффициент  $w_{ij}^{km}$  равен единице, если для  $j$ -ой записи  $i$ -го вида присутствует информация в медицинских картах пациентов  $k$  и  $m$ ; и равен нулю, если хотя бы у одного пациента  $j$ -ая запись  $i$ -го вида отсутствует. Например, пол и возраст в медкартах есть всегда, а результаты кардиограммы могут и не быть.

Обозначим через  $s_{ij}^{km}$  меру схожести пациентов  $k$  и  $m$  по  $j$ -ой записи  $i$ -го вида, которая для бинарных и порядковых данных вычисляется по формуле (1), а для количественных (непрерывных) данных – по формуле (2):

$$s_{ij}^{km} = \begin{cases} 1, r_{kij} = r_{mij} (k, m \in P), \\ 0, r_{kij} \neq r_{mij} (k, m \in P). \end{cases} \quad (1)$$

$$s_{ij}^{km} = 1 - \frac{|r_{kij} - r_{mij}|}{R_{ij}}, R_{ij} = \max_{p \in P} \{r_{pij}\} - \min_{p \in P} \{r_{pij}\}. \quad (2)$$

Тогда  $s^{km}$ , мера схожести между двумя медкартами пациентов  $k$  и  $m$ , вычисляется с помощью коэффициента Гауэра:

$$s^{km} = \frac{\sum_{i \in I} \sum_{j \in J_i} w_{ij}^{km} s_{ij}^{km}}{\sum_{i \in I} \sum_{j \in J_i} w_{ij}^{km}} \quad (3)$$

Полученная мера схожести сравнивается с заранее заданным пороговым значением. Результатом данного сравнения является группа медкарт, которые схожи по записям, представленным в них, с медкартой пациента, для которого производится оценка рисков развития заболеваний. Таким образом, обозначим через  $P' (P' \subset P)$  – множество медкарт пациентов, для которых мера схожести выше заданного порогового значения.

Для дополнения недостающих данных в медкарте  $k$ -го пациента необходимо вычислить оценки  $j$ -ой записи  $i$ -го вида, используя средневзвешенное значение полученных мер схожести:

$$r_{kij} = \frac{\sum_{m \in P'} S^{km} r_{mij}}{\sum_{m \in P'} S^{km}} \quad (4)$$

Таким образом можно дополнить недостающими данными медкарты пациентов, а затем полученную информацию использовать для оценки общего состояния пациента.

**Выводы.** Проведенные исследования доказали целесообразность использования техники КФ для нахождения рисков развития заболеваний. Но обработка медицинских данных представляет собой сложный процесс, так как данные характеризуются разнородностью и неполнотой, поэтому рассмотренная задача остается все еще актуальной.

**Список литературы:** 1. A Guide to the Project Management Body of Knowledge (PMBOK Guide). Режим доступа: [<http://www.pmi.org> 15.11.14]. 2. Daniel Lemire. Slope One Predictors for Online Rating-Based Collaborative Filtering / Daniel Lemire, Anna Maclachlan. Режим доступа: [[http://lemire.me/fr/documents/publications/lemiremaclachlan\\_sdm05.pdf](http://lemire.me/fr/documents/publications/lemiremaclachlan_sdm05.pdf) 19.11.14] 3. M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the Association of Computing Machinery*, 40(3):66-72,1997. 4. Мельник К.В. Применение аппарата Байесовых сетей при обработке данных из медицинских карточек / Мельник К.В., Глушко В.Н. // *Science and Education a New Dimension: Natural and Technical Sciences*. –I(2), Issue:15, 2013. – P.126–129. Венгрия, Будапешт. 5. Мельник К.В. Анализ данных для медицинской информационной системы в лечебно-профилактическом учреждении / Мельник К.В., Голоскоков А.Е. // *Вестник НТУ«ХПИ»*. – Харьков. – № 29. – 2012. – с.60–67. 6. Brian S. Everitt. *Cluster Analysis* / Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl // John Wiley & Sons, Ltd , 2001. – 330 p.

**Bibliography (transliterated):** 1. A Guide to the Project Management Body of Knowledge (PMBOK Guide). Web. 15 November 2014 <<http://www.pmi.org>>. 2. Daniel Lemire. Slope One Predictors for Online Rating-Based Collaborative Filtering / Daniel Lemire, Anna Maclachlan. Web. 19 November 2014 <[http://lemire.me/fr/documents/publications/lemiremaclachlan\\_sdm05.pdf](http://lemire.me/fr/documents/publications/lemiremaclachlan_sdm05.pdf)>. 3. M. Balabanovic and Y. Shoham. "Content-based, collaborative recommendation". *Communications of the Association of Computing Machinery*, No. 40(3). 66–72. 1997. Print. 4. Mel'nik, K. V., and V. N. Glushko. "Primeneniye apparata Bayesovykh setey pri obrabotke dannykh iz meditsinskikh kartochek" *Science and Education a New Dimension: Natural and Technical Sciences*. –I(2), No.15, Hungary, Budapest. 2013. 126–129. Print. 5. Mel'nik, K.V., and A.Ye. Goloskokov. "Analiz dannykh dlya meditsinskoy informatsionnoy sistemy v lechebno-profilakticheskom uchrezhdenii". *Vestnik NTU"KHPI"*. No. 29. Kharkiv. 2012. 60–67. Print. 6. Brian, S, et al. Everitt. *Cluster Analysis*. John Wiley & Sons, Ltd , 2001. Print.

*Поступила (received) 20.11.2014*