УДК 004.822

РАЗРАБОТКА СИСТЕМЫ ФОРМИРОВАНИЯ РЕКОМЕНДАЦИЙ ДЛЯ ПОЛЬЗОВАТЕЛЕЙ КОРПОРАТИВНОГО ПОРТАЛА УНИВЕРСИТЕТА

Ефимов М.Н., Шлей М.Д., Варенников Д.А.

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, e-mail: efimov@niuitmo.ru

Данная статья посвящена вопросам формирования индивидуальных рекомендаций для пользователей корпоративного портала информационной системы Университета ИТМО. Под рекомендациями в рамках данной статьи понимаются предложения к просмотру определенных информационных объектов: новых публикаций, конкурсов, проектов, тематика которых связана с интересами пользователя. Для определения рекомендаций предложен метод, основанный на понятии совместной фильтрации информации. Суть данного метода заключается в анализе информации о научных интересах, указанных пользователем, статистики по работе с информацией в корпоративном портале, активности пользователей имеющих схожие научные интересы. По результатам анализа формируется ранжированный список информационных объектов, которые могут быть интересны пользователю. Предложенные методы были реализованы в программной системе, которая была апробирована и внедрена в информационную систему университета.

Ключевые слова: совместная фильтрация, информационные системы, ранжирование информации, алгоритмы, проектная деятельность, базы данных

RECOMMENDATIONS SYSTEM DEVELOPMENT FOR USERS OF THE UNIVERSITY CORPORATE PORTAL

Efimov M.N., Shley M.D., Varenikov D.A.

St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Sankt-Peterburg, e-mail: efimov@niuitmo.ru

This article deals with the formation of individual recommendations for users of corporate information portal of ITMO University. The recommendations in this article refers to proposals to viewing some information objects: new publications, competitions, projects, topics which are related to the interests of the user. To determine the recommendations proposed method based on the concept of collaborative filtering information. The essence of this method lies in the analysis of information on research interests, user-specified, statistics on the information in a corporate portal, user activity have similar research interests. The analysis generated a ranked list of data objects that may be of interest to the user. The proposed methods have been implemented in a software system that has been tested and implemented in the information system of the university.

Keywords: collaborative filtering, information systems, information rankings, algorithms, project activity, database

На протяжении нескольких лет в Университете ИТМО функционирует информационная система, с которой работают все сотрудники и обучающиеся университета. Пользователи работают с разнообразной информацией, касающейся всех сфер деятельности: от индивидуальных планов аспирантов до ведения финансовой отчетности. Также в университете имеются проектные менеджеры, которые регулярно отслеживают новости об открываемых конкурсах и фондах и отбирают те из них, которые могут заинтересовать сотрудников и студентов университета [4].

В данной статье рассмотрены вопросы создания системы, позволяющей помочь проектным менеджерам распространить информацию об источниках финансирования научной деятельности среди пользователей корпоративного портала, а также информировать пользователей о появлении возможно интересных для него публикаций и проектов. Данная система включает в себя модуль оценки активности пользователей и модуль формирования списка рекомендованных к просмотру объектов. Предложенное решение основывается на методах совместной фильтрации с использованием метрик оценки схожести объектов фильтрации. Данные методы получили в последнее время широкое распространение в открытых поисковых системах в сети Интернет. В качестве площадки для реализации используется подсистема «ИСУ Портфолио», входящая в состав информационной системы университета [1, 2], в которой пользователь указывает свои научные интересы, контактную информацию, свои научные достижения, работает со своим индивидуальным планом. На основании внесенной пользователем информации программный модуль формирует список информационных объектов, которые могут быть интересны для него.

Общая схема работы системы представлена на рис. 1. Два пользователя работают с информационной системой. У них есть указанные ими сведения об их научных интересах, а также набор данных об их активности. Из этой информации мы можем получить набор ключевых слов для первого пользователя и для второго, на рисунке данные наборы изображены в виде кругов. Эти два круга являются множествами ключевых слов, которые пересекаются между собой в некоторой области. Ключевые слова, содержащиеся в кругах, имеют связи с различными информационными объектами. Следовательно, обладая сведениями о том, что множества ключевых слов пересекаются, можно сделать вывод, что информационные объекты, полученные из области интересов первого пользователя, будут интересны второму пользователю, а второго – первому. Размеры кругов и размер их пересечения характеризует схожесть интересов пользователей между собой [1].

В основе описанных в дальнейшем алгоритмов совместной фильтрации данных лежат сведения о схожести объектов. В качестве объектов в нашем случае могут выступать ключевые слова либо пользователи корпоративного портала. И те, и другие связаны друг с другом либо через непосредственное указание пользователем своих научных интересов, либо через данные об активности пользователя. Указать свои научные интересы (ключевые слова), пользователь может в подсистеме «ИСУ Портфолио».

Для накопления статистики о просмотрах пользователем информации о конкурсах, публикациях, проектах разработан соответствующий модуль. С помощью данного модуля накоплена информация об активности пользователей за 2015 год [2].

Таким образом, используя методы совместной фильтрации, можно реализовать возможность предоставления пользователям информационной системы университета списка информационных объектов (конкурсы, проекты, публикации), рекомендованных к просмотру.

Реализуемые методы совместной фильтрации основываются на понятии схожести объектов [3], которая заключается в том, что объекты имеют схожий набор характеристик (ключевых слов). И при просмотре одних объектов пользователя могут заинтересовать другие объекты, схожие по характеристикам. На основании оценки схожести объектов происходит их ранжирование и, как следствие, выдача рекомендаций.

Будем считать, что связь пользователя с ключевым словом при непосредственном указании связи человеком весомее, чем связь с ключевым словом, полученная из его активности. Данное предположение обосновывается тем, что научный интерес указанный человеком однозначно релевантен по отношению к человеку, иначе он бы его не стал указывать, а активность пользователя может основываться на далеких от интересов человека факторах. Человек мог зайти на страницу случайно, или ему стал любопытен какой-то несущественный момент.

Для определения оценки схожести используются различные метрики. В ходе разработки системы рассматривались следующие метрики: Евклидово расстояние, коэффициент Пирсона, коэффициент Жаккара.

В случае Евклидова расстояния: пусть A – множество ключевых слов, с которыми имеет связь первый пользователь, а B – множество ключевых слов, с которыми имеет связь второй пользователь, тогда множество C будет равно объединению множеств A и B:

$$C = A \cup B. \quad (1)$$

Тогда на основании выше сказанного коэффициент схожести между двумя пользователями можно рассчитать как

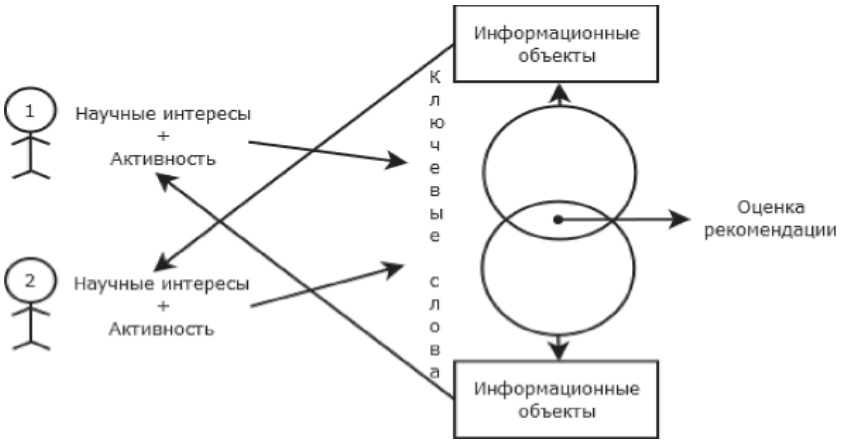
$$K = \sqrt{\sum_i (\omega_{i1} \cdot x_{i1} - \omega_{i2} \cdot x_{i2})^2}, \quad (2)$$

где ω_{ij} – вес связи пользователя j с ключевым словом i из множества C (в рамках данной работы параметр может принимать два значения, которые определены эмпирически: для ключевых слов, определенных самим пользователем, $\omega = 1$, для ключевых слов определенных на основе его активности, $\omega = 0{,}5$); $x_{ij} \in \{0, 1\}$ – факт наличия связи пользователя j с ключевым словом i [3].

Воспользуемся формулой (2) для расчета коэффициента схожести между двумя пользователями. Коэффициент, вычисленный по этой формуле, будет тем меньше, чем больше сходства между людьми. Однако нам нужна формула, результат расчетов по которой тем больше, чем люди более похожи друг на друга. Этого можно добиться, добавив к вычисленному значению 1 (чтобы никогда не делить на 0) и взяв обратную величину:

$$K = \frac{1}{1 + \sqrt{\sum_i \left(\omega_{i1} \cdot x_{i1} - \omega_{i2} \cdot x_{i2}\right)^2}}. \quad (3)$$

Коэффициент, вычисленный по данной формуле, принимает значение от 0 до 1, причем 1 получается, когда предпочтения двух людей в точности совпадают.



Общая схема совместной фильтрации в информационной системе университета

Для коэффициента Пирсона: коэффициент схожести между двумя пользователями можно рассчитать как

$$K = \frac{\sum_i \omega_{i1} \cdot x_{i1} \cdot \omega_{i2} \cdot x_{i2} - \frac{\sum_i \omega_{i1} \cdot x_{i1} \sum_i \omega_{i2} \cdot x_{i2}}{n}}{\sqrt{\left(\sum_i (\omega_{i1} \cdot x_{i1})^2 - \frac{\left(\sum_i \omega_{i1} \cdot x_{i1}\right)^2}{n}\right)\left(\sum_i (\omega_{i2} \cdot x_{i2})^2 - \frac{\left(\sum_i \omega_{i2} \cdot x_{i2}\right)^2}{n}\right)}}, \quad (4)$$

где ω_{ij} – вес связи пользователя j с ключевым словом i из множества C; $x_{ij} \in \{0, 1\}$ – факт наличия связи пользователя j с ключевым словом i; n – количество элементов в множестве C.

Множество C определяется формулой (1).

Воспользуемся формулой (4) для расчета коэффициента корреляции Пирсона между людьми. Результат вычисления в общем случае может лежать от –1 до 1. В случае идеальной корреляции коэффициент принимает значение 1, а в случае идеальной обратной корреляции – значение –1. При отсутствии корреляции коэффициент принимает значение 0.

Для вычисления схожести двух пользователей между собой с помощью коэффициента Жаккара воспользуемся следующей формулой:

$$K = \frac{c}{a + b - c}, \quad (5)$$

где c – количество ключевых слов, с которыми связаны и первый, и второй пользователи; a – количество ключевых слов, с которыми связан только первый пользователь; b – количество ключевых слов, с которыми связан только второй пользователь.

Для каждой метрики был разработан алгоритм по расчету оценки сходства объектов, также было проведено сравнение результатов, получаемых в ходе их работы. Для этого была взята выборка из 100 пользователей, заполнивших свои личные профили и имеющих высокую активность по работе с системой университета. Между всеми пользователями была рассчитана схожесть по научным интересам и определено время расчета всех оценок схожести. Тестирование проводилось на сервере Sun SPARC Enterprise M4000 с процессором SPARC64 VII+.

Результаты сравнения метрик по быстродействию

Используемая метрика	Время работы алгоритма, секунды
Евклидово расстояние	11,765
Коэффициент корреляции Пирсона	12,265
Коэффициент Жаккара	12,547

На основании проведенных экспериментов в качестве основной метрики, используемой в дальнейшем, был выбран коэффициент Жаккара. Его преимуществом является более качественное определение пар пользователей имеющих схожие научные интересы. Данное соответствие было

получено на основе экспертной оценки в результате анализа творческих коллективов сотрудников, совместных публикаций и выполняемых проектов. При этом производительность данного алгоритма не сильно отличается от двух других.

Рассмотрим объекты, которые описываются множеством равноценных по отношению к объекту ключевых слов. Тогда будем рекомендовать те объекты, которые имеют связь с ключевым словом присутствующим в ранжированном списке, полученном ранее. При этом получится весьма громоздкий список объектов.

Полученный список необходимо отсортировать. Данная процедура не имеет однозначной реализации. В качестве одной из возможных реализаций можно предложить сортировать объекты по одному ключевому слову, имеющему максимальный вес в выдаче, полученной ранее. В случае равенства весов или одинаковых ключевых слов у объектов – объекты сортируются по дополнительным параметрам, индивидуальным для каждого типа объектов.

Такой подход к сортировке данных позволяет избежать проблем, которые бы возникали при нахождении среднего веса объекта по всем его ключевым словам. Это могло бы обесценить связь с релевантным ключевым словом, в случае если объект связан с большим количеством нерелевантных ключевых слов.

Однако такой подход имеет и свои недостатки. Ключевое слово может быть связано с объектом очень слабой логической связью, однако все связи формально имеют одинаковый вес. В таком случае пользователь получит нерелевантную выдачу.

Полученный механизм рекомендации выстроен таким образом, что для создания набора данных необходима информация обо всех связях пользователей с ключевыми словами. На небольшом наборе данных данный подход показывает хорошие результаты, но при большом количестве пользователей сравнение каждого пользователя со всеми другими, а затем сравнение ключевых слов, которые связаны с каждым пользователем займет недопустимо много времени. Помимо этого, на данный момент имеющиеся данные довольно сильно разряжены, поэтому целесообразным становится другой подход.

То, что было описано ранее, носит название совместной фильтрации по схожести пользователей. Альтернатива известна под названием совместная фильтрация по схожести образцов. Когда набор данных достаточно большой, то совместная фильтрация по схожести образцов может давать лучшие результаты, причем многие вычисления можно выполнить заранее, поэтому пользователь получит рекомендации быстрее.

Процедура фильтрации по схожести образцов во многом основана на вышеизложенном материале. Основная идея заключается в том, что для каждого образца заранее вычислить большинство схожих на него объектов [5]. Тогда для выработки рекомендаций пользователю достаточно будет найти те ключевые слова, с которыми он связан, и составить взвешенный список ключевых слов, максимально похожих на связанные с пользователем. Хотя на первом шаге необходимо исследовать все данные, результаты сравнения образцов изменяются не так часто, как результаты сравнения пользователей. Это означает, что не нужно постоянно пересчитывать для каждого образца список похожих на него. Пересчет списка образцов можно организовать ночью, раз в сутки.

Предложенный подход реализован в виде алгоритма определения рекомендаций ключевых слов для пользователя. На вход алгоритма подается информация обо всех пользователях системы в виде массива объектов P_i, $i = 1...n$. В результате работы алгоритма получается ранжированный список ключевых слов для каждого пользователя, которые могут быть ему интересны. Данный алгоритм можно описать следующими шагами:

Шаг 1. Определяется круг интересов для каждого пользователя P_i в виде набора ключевых слов из его профиля и анализа активности (множества A_i).

Шаг 2. Рассчитываются оценки схожести интересов между всеми пользователями при помощи метрики Жаккара (формула (5)). Получаем матрицу оценок схожести $K_{i,j}$, $i = 1...n, j = 1...n$.

Шаг 3. Для каждого пользователя выполняем ранжирование списка других пользователей в соответствии с полученными значениями. От большего значения к меньшему.

Шаг 4. В соответствии с порядком пользователей им сопоставляются их интересы и ранжируются в том же порядке. В случае наличия ключевого слова у двух разных пользователей с различным показателем схожести – приоритет отдается пользователю с наибольшим коэффициентом. То есть ключевое слово занимает наиболее высокую позицию при ранжировании.

Шаг 5. Из полученных списков выбираются первые D ключевых слов (D задается администратором системы, в ходе исследования использовалось $D = 10$). В итоге для каждого пользователя формируется ранжированный список (группа) ключевых слов $R_{i,j}$, $i = 1...n, j = 1...D$.

Шаг 6. Далее для каждого ключевого слова определяется набор ключевых слов, которые наиболее часто встречаются с ним в других группах (при этом должно выполняться условие встречи более чем в 90% групп). В результате для каждого ключевого слова формируется набор связанных с ним ключевых слов – образцов схожести.

Шаг 7. На основании определенного для пользователя круга интересов (множество A_i) и построенных образцов схожести формируется список рекомендованных для пользователя ключевых слов.

Полученный в ходе работы алгоритма набор ключевых слов (совместно с ключевыми словами, указанными в его научных интересах) в дальнейшем используется для рекомендации пользователю информационной системы университета просмотра информации о конкурсах, грантах, публикациях и проектах. Дальнейшая работа по данной тематике будет направлена на поиск схожих научных интересов у пользователей информационной системы университета и пользователей открытых научных Интернет-ресурсов [6].

Предложенный алгоритм был реализован в системе определения рекомендаций. Данное решение было успешно апробировано в рамках корпоративного портала университета. Пользователи получили удобный инструмент для доступа к интересующей их информации. Разработанная система была зарегистрирована в государственном фонде РОСПАТЕНТ (свидетельство о регистрации № 2015616767 от 22 июня 2015 г.).

Список литературы

1. Ефимов М.Н., Шлей М.Д., Варенников Д.А. Метод определения рекомендаций для пользователей информационной системы на основе их научных интересов и активности // Научно-образовательная информационная среда XXI века: сборник материалов конференции. – Петрозаводск, 2014. – С 71–73.

2. Ефимов М.Н., Шлей М.Д., Варенников Д.А. Система определения научных интересов пользователей // Телематика – 2014: Труды XXI Всероссийской научно-методической конференции. – СПб., 2014. – С. 87–88.

3. Иванов Е.Е., Шустов Д.А., Перешивкин С.А. Многомерные статистические методы [Электронный ресурс]. – 2010. – URL: http://ecocyb.narod.ru/513/MSM/msm2_2.htm (дата обращения: 17.04.2015).

4. Казин Ф.А., Биккулов А.С., Зленко А.Н., Тойвонен Н.Р., Попова И.А., Шлей М.Д., Варенников Д.А. Система поддержки проектной деятельности в Университете ИТМО // Инновации. – 2014. – № 8(190). – С. 77–83.

5. Сегаран Т. Программируем коллективный разум. – СПб.: Символ-Плюс, 2008. – 368 с.

6. Семерханов И.А., Муромцев Д.И. Интеграция информационных систем на основе технологии связанных данных // Научно-технический вестник информационных технологий, механики и оптики. – 2013. – № 5 (87). – С. 123–127.

References

1. Efimov M.N, Shley M.D., Varenikov D.A. *Sbornik materialov konferentsii* «*Nauchno-obrazovatelnaya informatsionnaya sreda XXI veka*» (Proceedings of the conference «Scientific and educational information environment of the XXI century»), Petrozavodsk, 2014, pp. 71–73.

2. Efimov M.N., Shley M.N., Varenikov D.A. Trudy XXI *Vserossiyskoy nauchno-metodicheskoy konferentsii* «*Telematika–2014*» (Proceedings of the XII All-Russian Scientific Conference «Telematics 2014»), Sankt-Peterburg, 2014, pp. 87–88.

3. Ivanov E.E., Shustov D.A., Pereshivkin S.A. *Mnogomernye statisticheskie metody* (Multivariate statistical methods), Available at: http://ecocyb.narod.ru/513/MSM/msm2_2.htm (accessed 17 April 2015).

4. Kazin F.A., Bikkulov A.S., Zlenko A.N., Toyvonen N.R., Popova I.A., Shley M.D., Varenikov D.A. *Innovatsii*, 2014, no. 8(190), pp. 77–83.

5. Segaran T. *Programmiruem kollektivnyy razum* (Programming Collective Intelligence). Sankt-Peterburg, 2008, 368 p.

6. Semerhanov I.A., Muromtsev D.I. Nauchno-tehnicheskiy vestnik informatsionnyh tehnologiy, mehaniki i optiki, 2013, no. 5(87), pp. 123–127.