Recommender systems "Entrant" for reception of high school committee

Mityanina A.[1], Gazha K.[2]

Рекомендательная система «Абитуриент» для приемной комиссии вуза

Гажа К. В.[1], Митянина А. В.[2]

[1]Митянина Анастасия Владимировна / Mityanina Anastasiya Vladimirovna – преподаватель;
[2]Гажа Константин Владимирович / Gazha Konstantin Vladimirovich – студент,
кафедра информационных технологий и экономической информатики,
Институт информационных технологий,
Челябинский государственный университет, г. Челябинск

Аннотация: *в статье описывается концепт рекомендательной системы для выполнения качественного набора абитуриентов во время приемной кампании Института информационных технологий Челябинского государственного университета. Она будет прогнозировать вероятности поступления абитуриентов на выбранные ими направления обучения и классифицировать их на определенные группы с помощью математико-статистических методов анализа данных. Члены приемной комиссии смогут акцентировать своё внимание на наиболее предпочтительной группе абитуриентов. В статье также приводится обзор аналогов данной системы.*

Abstract: *in this article describes a concept of a recommendation system for Institute of information technologies of Chelyabinsk State University, which could help to manage their recruitment efforts much better. It will predict the probabilities of school-leavers' admission in order to their chosen directions of study and classify them into certain groups with the help of mathematical and statistical methods of data analysis. Members of the admissions office will be able to focus its attention on the most preferred group of matriculants. There is a review of analogues in this article.*

Ключевые слова: *анализ данных, приемная кампания, абитуриенты, классификация.*
Keywords: *data-mining, admissions office, school-leavers, classification.*

В Российской Федерации государственные университеты предоставляют возможность студентам учиться на бюджетной основе. Как правило, количество абитуриентов, желающих поступить на бюджетные места обучения, гораздо больше числа выделяемых мест, из-за чего существует конкурс среди абитуриентов.

В соответствии с законом об образовании РФ, абитуриент имеет право подать наравне с оригиналом копии своих документов, но не более чем в пять вузов. В связи с этим приемная комиссия университета не может знать наверняка, какой из вузов в конечном итоге выберет абитуриент. Более того, в рамках одного университета абитуриент может выбрать не более трёх направлений обучения. При подаче заявления на поступление, абитуриент определяет приоритеты направлений, в которых он заинтересован. В случае если он выигрывает конкурс на бюджетное место и желает учиться в этом университете, абитуриент подает оригиналы документов.

Абитуриенты, которые набрали большой суммарный балл, с большой вероятностью попытаются поступить в более престижные университеты страны, а для подстраховки подадут копии документов ещё в несколько дополнительных вузов. Этот случай хорошо описывает группу абитуриентов с высоким уровнем подготовки.

Также можно выделить группу людей, которые набрали не столь высокие баллы. Если становится ясно, что представитель этой группы не проходит по баллам ни на одно из направлений обучения в предпочтительном для него вузе, а в другом университете он может занять высокую позицию в списке претендентов на бюджетное место, то, скорее всего, он отдаст предпочтение тому вузу, где сможет учиться на бюджете.

Сформулируем проблему для Института информационных технологий ЧелГУ во время приемной кампании: сделать качественный набор в соответствии с планом приема, учитывая тот факт, что абитуриенты ИИТ ЧелГУ могут в конечном итоге отдать предпочтение другому университету. Для решения данной проблемы будет проведено научное исследование.

Основная цель исследования – найти зависимость между информацией об абитуриенте и выбранным им местом обучения. Иначе говоря, классифицировать абитуриентов согласно их признакам и конечным решениям, в какой вуз они поступят. В зависимости от класса, в который попал абитуриент, со стороны приемной комиссии в отношении этого абитуриента будет определяться совокупность действий: рассылка персональных писем, звонки. В ходе исследования будет спроектирована и реализована система, которая позволит оценивать вероятность поступления того или иного абитуриента в ИИТ ЧелГУ, и вследствие этого давать рекомендации по взаимодействию с поступающим. Таким образом, объектом

исследования является приемная кампания ИИТ ЧелГУ. Предмет исследования – классификация абитуриентов бакалавриата.

Анализ аналогов

Приемная комиссия колледжа Covenant применяет методы анализа данных, чтобы определить, с какой вероятностью конкретный абитуриент поступит к ним на обучение, основываясь на его месте жительства, интересах и способностях к обучению [1]. Приемная комиссия фокусирует свое внимание на группе абитуриентов с высоким показателем вероятности поступления, делает рассылку персональных писем. Из-за специфики зарубежного образования, в данном решении не используется информация о позиции конкретного абитуриента в рейтингах других вузов.

В статье «Generalized Net Model of Using Data Mining Techniques for Process of Undergraduate Matriculation in a Digital University» описано исследование, посвященное созданию системы для определения вероятности поступления студента в университет на дистанционное обучение [2]. Для достижения этой цели использовались обобщенные сети (модификация сетей Петри). В статье не демонстрируются результаты работы данного решения. Не обоснован выбор входных данных для анализа.

Факультет математики и информатики Гродненского государственного университета имени Янки Купалы рассылал определенной группе абитуриентов (наиболее предпочтительных) письма-приглашения о поступлении [3]. Для оценки экономической эффективности данного метода привлечения абитуриентов, было проведено исследование, в котором применялся метод бутстреп-анализа. В результате получили, что из общего числа зачисленных абитуриентов, 61 получили письма-приглашения, 15 из которых поступило на факультет исключительно благодаря рассылке. Данное исследование прекрасно демонстрирует, насколько эффектно взаимодействие с группой абитуриентов. Исходя из содержания статьи непонятно, как выделялась группа потенциальных кандидатов на поступление, в исследовании не проводилась классификация абитуриентов (не давалась вероятностная оценка поступления того или иного абитуриента в вуз).

В статье «Качество образования: data-mining баз данных результатов центрального тестирования» описано исследование, которое проводилось с целью выявить целевую аудиторию из потока абитуриентов одного из вузов Республики Беларусь [4]. В ходе кластерного анализа получили, что данные, состоящие из сумм баллов абитуриента за центральное тестирование, находятся очень далеко от данных, содержащих результаты сессии. В этом исследовании не использовались данные об абитуриентах в других вузах.

В китайском вузе «Jiangsu University of Science and Technology» проводилось исследование с целью выполнения плана приема путём создания рекомендательной системы для абитуриентов с высшим образованием, желающих продолжить своё обучение в магистратуре или аспирантуре [5]. Система принимает на вход характеристики абитуриента и в результате анализа выдает наиболее подходящие вузы и направления для обучения. Для решения этой задачи использовался алгоритм ID3.

Все вышеперечисленные аналоги являются некоммерческими продуктами, отсюда следует высокая практическая значимость данной научно-исследовательской работы.

Концепция системы. Сбор данных

Чтобы спрогнозировать, куда поступит в конечном итоге абитуриент, подавший документа в ИИТ ЧелГУ, будет разработан алгоритм классификации. Для его реализации необходимо иметь матрицу объектов-признаков и соответствующую ей матрицу ответов. В матрице объектов-признаков строки соответствуют объектам (абитуриентам), а столбцы их признакам (определенным характеристикам). Матрица ответов имеет размерность $1 \times n$, где n – число объектов. В ней содержится информация о месте, в которое поступил соответствующий абитуриент (вуз, факультет, направление).

На данный момент не определены все признаки, которые будут задействованы в анализе данных. Но уже сейчас можно привести в пример такие как пол, место жительства абитуриента, его баллы за ЕГЭ.

Первичные данные об абитуриентах будут собираться из пакета документов, необходимых для поступления в вуз. Эта информация впоследствии будет занесена в базу данных разрабатываемой системы.

Для получения сведений о том, подал ли абитуриент документы в другие вузы, будут использоваться специальные поисковые программы – краулеры. С их помощью будет проводиться сбор информации об абитуриентах с официальных сайтов других вузов. В случае обнаружения в рейтинговых списках этих вузов человека, который претендует учиться в ИИТ ЧелГУ, программа сохранит в базу данных определенный набор информации: факт участия абитуриента в конкурсе на бюджетные места, его позиция в рейтинге, число бюджетных мест, тип документа об образовании (оригинал или копия).

Рекомендательная система разрабатывается для Института информационных технологий ЧелГУ, следовательно, в анализе данных нужно учесть как можно больше вузов из Уральского федерального

округа. В то же время, ИИТ готовит будущих специалистов в технической сфере, поэтому нецелесообразно учитывать вузы медицинской, гуманитарной направленности. Также в список университетов для анализа данных будут включены престижные вузы страны.

В ходе первой приемной кампании будут получены результаты о том, куда в конечном итоге поступил тот или иной абитуриент, подавший документы в ИИТ ЧелГУ. Таким образом, будет известна вся необходимая информация для начала разработки классификатора. За основу будут взяты одни из распространенных алгоритмов: метод решающих деревьев, метод ближайших соседей, линейные классификаторы, нейронные сети. По результатам сравнительного анализа результатов работы алгоритмов получим наиболее эффективные из них. Под эффективностью подразумевается точность алгоритма – одна из метрик качества классификации.

Также важной задачей исследования является поиск значимых признаков. Не все характеристики абитуриента, которые используются в анализе данных, могут быть полезны. Некоторые признаки могут снизить качество классификации, следовательно, от них нужно избавляться и не учитывать их в дальнейшем.

После проведения первой приемной кампании будет разработан алгоритм классификации абитуриентов, будут определены признаки абитуриентов, которые оказывают существенное влияние на результат классификации.

Способы и технологии реализации

В результате будет разработана рекомендательная система для приемной комиссии (Рисунок 1).

Рис. 1. Архитектура системы

Далее в статье приводится предварительный выбор инструментария.

Веб-сайт для приемной комиссии будет написан на языке C# с использованием фреймворка APS.NET. Данный выбор обусловлен высокой надежностью данной платформы, расширяемостью.

В качестве базы данных для хранения информации об абитуриентах планируется использовать PostgreSQL, которая ничем не хуже аналогов, вдобавок распространяется бесплатно.

Краулеры для сбора данных об абитуриентах с сайтов других вузов и скрипты для анализа данных будут реализованы на языке Python, так как для него существует множество вспомогательных библиотек для достижения цели исследования.

Литература

1. *Christopher J. S.* A data mining study of the matriculation of Covenant college applicant // Proceedings of the 46th Annual Southeast Regional Conference on XX (Оберн, США, 28 марта 2008). Нью-Йорк: Изд-во ACM SE, 2008. С. 209-214.
2. *Anthony S., Evdokia S., Daniela O.* Generalized Net Model of Using Data Mining Techniques for Process of Undergraduate Matriculation in a Digital University // International Workshop on Generalized Nets (София, Болгария, 5 декабря 2010). София: Изд-во Conference proceedings, 2010. С. 1-6.
3. *Петров С. В., Балдин Е. М., Лявшук В. Е.* Анализ данных с R // Linux Format/ 2010. № 2. С. 3-11.
4. *Ровба Е. А., Бойко В. К., Войтукевич Ю. А., Лявшук В. Е., Петров С. В.* Качество образования: data-mining баз данных результатов централизованного тестирования // Университетское управление: практика и анализ, 2012. № 5. С. 78-87.

5. *Qu Ya Hui*. Research on the Application of Data Mining Technology in the Regulating Matriculation for Postgraduate. [Электронный ресурс]: Аннотация к статье. URL: http://ysidata.com/showinfo-58-60212-0.html (дата обращения: 19.05.16).