Санкт-Петербургский Государственный Университет

Математико-механический факультет

Кафедра системного программирования

Разработка и реализация метода выявления зависимостей между образовательными материалами на основании статистики использования платформы онлайн-обучения Розалинд

Курсовая работы студента 445 группы

Колобова Романа Евгеньевича

Научный руководитель: Вяххи Н. И.

Санкт-Петербург

2013

Оглавление

Введение

EDM (educational data mining) – дисциплина, занимающаяся выделением закономерностей из данных большого объёма, полученных из образовательных сред. От классического анализа данных (data mining) она отличается в первую очередь тем, что использует методы и оценки, характерные для психометрии (теория и методика психологических измерений, в т. ч. оценки знаний) и анализ с помощью моделей (discovery with models).

Проблемы, встающие перед исследователями в этой области, весьма разнообразны. Так, например, одной из самых популярных тем является моделирование для предсказания поведения обучающихся и для определения наиболее выгодных стратегий обучения ([1], [2]). Для этих целей используются методы машинного обучения с учителем и без учителя (supervised and unsupervised classification) [3], построение деревьев принятия решений (decision trees) [4], скрытые марковские модели [5]. Классификация и кластеризация также привлекаются для определения уровня сложности задач [6]. Похожие методы применяются для выделения ошибочных представлений (концепций) об области знания [7]. Причём спектр целевых обучающих сред весьма широк: от простых наборов тематически связанных тестов [7] до «исследовательских» обучающих систем (exploratory learning environments), делающих ставку на свободное исследование пользователем интерактивных учебных материалов [3].

Таким образом, эта молодая и бурно развивающаяся область на сегодняшний день объединяет в себе множество различных методов для абсолютно непохожих между собой образовательных сред, и зачастую сам формат среды накладывает ограничения на использование многих уже разработанных методов, или, по крайней мере, заставляет прикладывать значительные усилия к их адаптации.

Проект Rosalind [8], использующий нестандартную в кругах онлайн-образования и привлекательную модель ориентированного графа для структурирования материала, представляет собой как раз один из таких недостаточно изученных и пока мало используемых форматов, который, тем не менее, хорошо подходит для технических дисциплин. Соответственно, встаёт задача подбора подходящих и формулирования новых методов для оптимизации взаимодействия пользователя и образовательной системы такого образца.

Постановка задачи

Rosalind – проект для онлайн-обучения биоинформатике, разрабатываемый в Санкт-Петербургском Академическом Университете совместно с University of California, San Diego. Также в рамках сотрудничества с проектом Coursera [9] в скором времени будет запущен онлайн-курс с лекционной частью на базе Coursera и с решением задач в Rosalind.

Ключевыми особенностями проекта на данный момент является древовидное построение учебного материала и упор на задачи, а не на короткие тесты и лекции – редко встречающаяся комбинация. По мере решения задач пользователю открывается доступ к новым, более сложным.

Главным приоритетом в разработке проекта такого рода является удобство обучающегося при работе со средой обучения; основной целью, соответственно – чтобы как можно больше пользователей продвинулось в обучении как можно дальше. Успешность отдельного обучающегося измерить несложно: мерой в данном случае является количество решённых задач и глубина прорешивания дерева. Измерить же, насколько оптимально организован обучающий материал в целом (или для данного конкретного пользователя) и как обеспечить наибольшую вовлечённость обучающегося – задача нетривиальная, и, как уже было сказано выше, в последнее время привлёкшая немалое количество исследователей.

Коротко говоря, нам хочется улучшить опыт взаимодействия студентов от общения с проектом. Глобально это означает, что необходимо предлагать пользователю самые удобные для него варианты – то есть быть максимально гибкими и полезными данному человеку. Для этого, естественно, нужен целый комплекс мер, а не один алгоритм. Но прежде чем вводить персонифицированные алгоритмы работы, надо устранить те изъяны в модели, которые отрицательно влияют на большинство пользователей.

Сейчас в базе проекта есть 128 задач и порядка 10000 пользователей, что уже позволяет делать выводы на основании собранных и обработанных данных.

На данный момент граф задач (ориентированный, без циклов; см. рис. 1) строится экспертным методом, с возможностью добавления новых задач «продвинутыми» пользователями. С ростом количества задач будет становиться всё сложнее оценивать

предварительные требования для каждой новой задачи (грубо говоря, искать входящие рёбра), и уже сейчас мы столкнулись с жалобами пользователей на чрезмерную сравнительную сложность некоторых предлагаемых для решения задач (т. е. после ряда легко решаемых задач открывается доступ к новой, гораздо более трудной, без смягчающих переходов). И хорошо, когда такие обратная связь есть, а ведь значительная часть пользователей по определённым причинам не хотят или не готовы давать отзывы, и могут просто молча разочароваться в проекте и уйти с него.

Поэтому одной из первоочерёдных задач стало автоматическое обнаружение наиболее трудных для пользователей переходов и возможных недостающих связей в дереве.

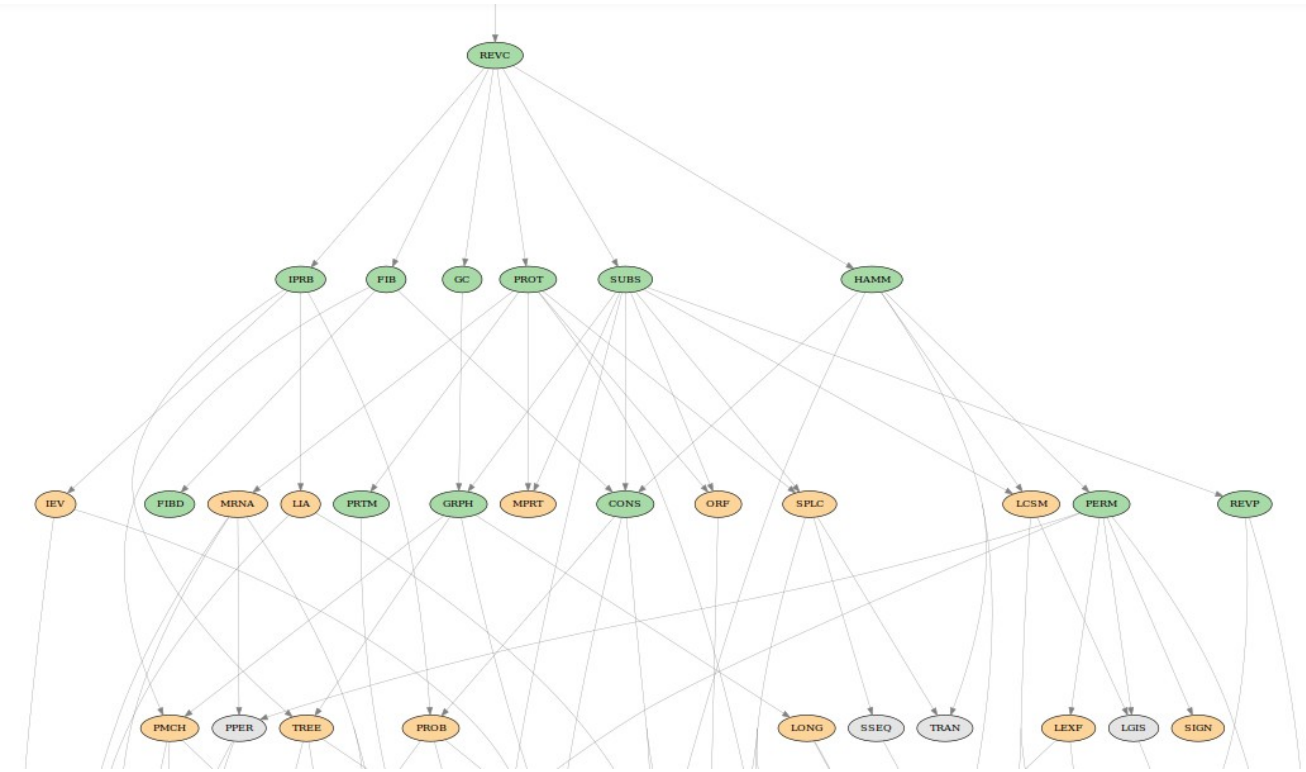

Рис. 1. Фрагмент графа задач Rosalind.

Зелёным цветом помечены решённые задачи, жёлтым – открытые, серым – недоступные.

Существующие подходы

Наиболее близкими к модели Rosalind являются проекты Khan Academy [10] и Knewton [11]. Большинство же других сред обучения используют линейную модель в подаче материала (например, подход Coursera с еженедельными наборами лекций и тестов, прогрессирующих в уровне сложности).

Khan Academy – ресурс для обучения предметам, в основе своей близким к школьной программе. В пределах предмета темы сгруппированы в граф, связи в котором носят скорее рекомендательный характер, в пределах темы же тесты можно проходить в свободном порядке. К сожалению, связанных с проектом статей или докладов по внутреннему устройству системы в сети найти не удалось, поэтому единственное, что мы можем почерпнуть из этого ресурса – набор возможностей, которыми можно было бы дополнить Rosalind, что к теме данной работы отношения не имеет.

Второй ресурс, Knewton – платформа для построения собственных курсов и работы с группами обучающихся, с богатыми возможностями по сбору и анализу статистики – прямо указывает на методы и оптимизации, использующиеся в системе [12], такие как Item Response Theory для более точной оценки тестов, графические вероятностные модели для составления рекомендаций, иерархическая кластеризация для разбиения студентов на группы обучения по уровню и подходу к пониманию материала. Структура графа же знаний (knowledge graph) проекта является, судя по всему, зафиксированной, в том смысле, что изменения влечёт за собой только добавление нового материала, а не статистика использования старого, поэтому и здесь мы не можем опереться на существующий опыт.

Вообще говоря, для определения зависимости между материалами обычно хватает экспертного метода, но в случае же Rosalind привлечь данный подход для оценки модели не позволяет ограниченность ресурсов проекта, сравнительная новизна области (биоинформатики) и отсутствие для неё устоявшихся практик обучения. И индикатором того, что стоит привлечь автоматический анализ, стали, как уже было выше сказано, негативные отзывы пользователей.

Реализация

Для работы был использован язык R, как наиболее подходящий для быстрого прототипирования алгоритмов машинного обучения и анализа данных (впоследствии полученные методы будут реализованы на Python).

Для начала, в рамках вспомогательного анализа графа задач, были реализованы простые оценки учебного материала, основанные на чисто статистическом подходе. Так, например, поскольку сайтом пользуются люди из разных стран, говорящие на разных языках, появилось подозрение, что некоторые задачи могут вызывать трудность исключительно из-за языкового барьера. Для этого было проведено сравнение соотношений долей русскоговорящих и англоговорящих пользователей, решивших задачу, для всех задач. Однако существенные расхождения обнаруживались только на задачах, решённых малым количеством пользователей, что обусловлено малым размером выборки и не является основанием для внесения изменений.

Следующей идеей было найти, какие задачи наиболее часто приводят к уходу или длительному отсутствию пользователей на проекте. Для этого для каждой задачи был посчитан коэффициент отмирания пользователей (аналог коэффициента текучести, churn rate) – доля пользователей, которые, не решив данную задачу, отсутствовали на проекте определённое количество времени (например, 2 недели). Однако, и в этом случае для различных опробованных периодов отсутствия существенно отличные доли были получены только на первых задачах, про которые с самого начала было очевидно, что они отсеивают большой процент пользователей, пришедших просто «попробовать» ресурс из любопытства (см. табл. 1).

Задача	Пользователей прочитало условие	Пользователей, долго отсутствовавших после этой задачи	Коэффициент отмирания
MULT	39	4	10.25%
LONG	352	43	12.21%
KMP	510	71	13.92%
GC	1902	296	15.56%
DNA	4714	1126	23.88%

Табл. 1. Коэффициенты отмирания пользователей для различных задач.

Затем было решено перейти к главной цели – нахождению связей между задачами, отсутствующих в текущем графе, но которые могли бы увеличить процент решаемости. В основном интересовали связи внутри уровней (граф разбит на слои, положение задачи в слое зависит от расстояния от корня и косвенно указывает на сложность задачи), т. е. те, которые указывали бы на большую сложность задачи, находящейся на конце ребра, по сравнению с другими задачами на её уровне.

Таким образом, было решено было найти найти те рёбра, которые связывали бы задачи, не находящиеся ещё в соотношении подчинения (в транзитивном замыкании исходного графа задач), и для которых выполнялось бы следующее условие: пусть мы рассматриваем ребро $A \to B$, A и B – задачи. Пусть a – количество пользователей, решивших задачу A и попробовавших после этого решить задачу B; ab– количество пользователей, решивших задачу A и после этого решивших задачу B; b – количество пользователей, попробовавших задачу B, не решив перед этим задачу A, и bb – количество пользователей, решивших задачу B, и не решивших перед этим задачу A. Тогда если $\frac{a}{ab} - \frac{b}{bb}$ больше некоего порогового значения, то будем считать ребро $A \to B$ «отсутствующим».

Данный подход был реализован, и на практике получились результаты, согласовывавшиеся с вышеуказанными жалобами пользователей на сложность определённой задачи (см. рис. 2).

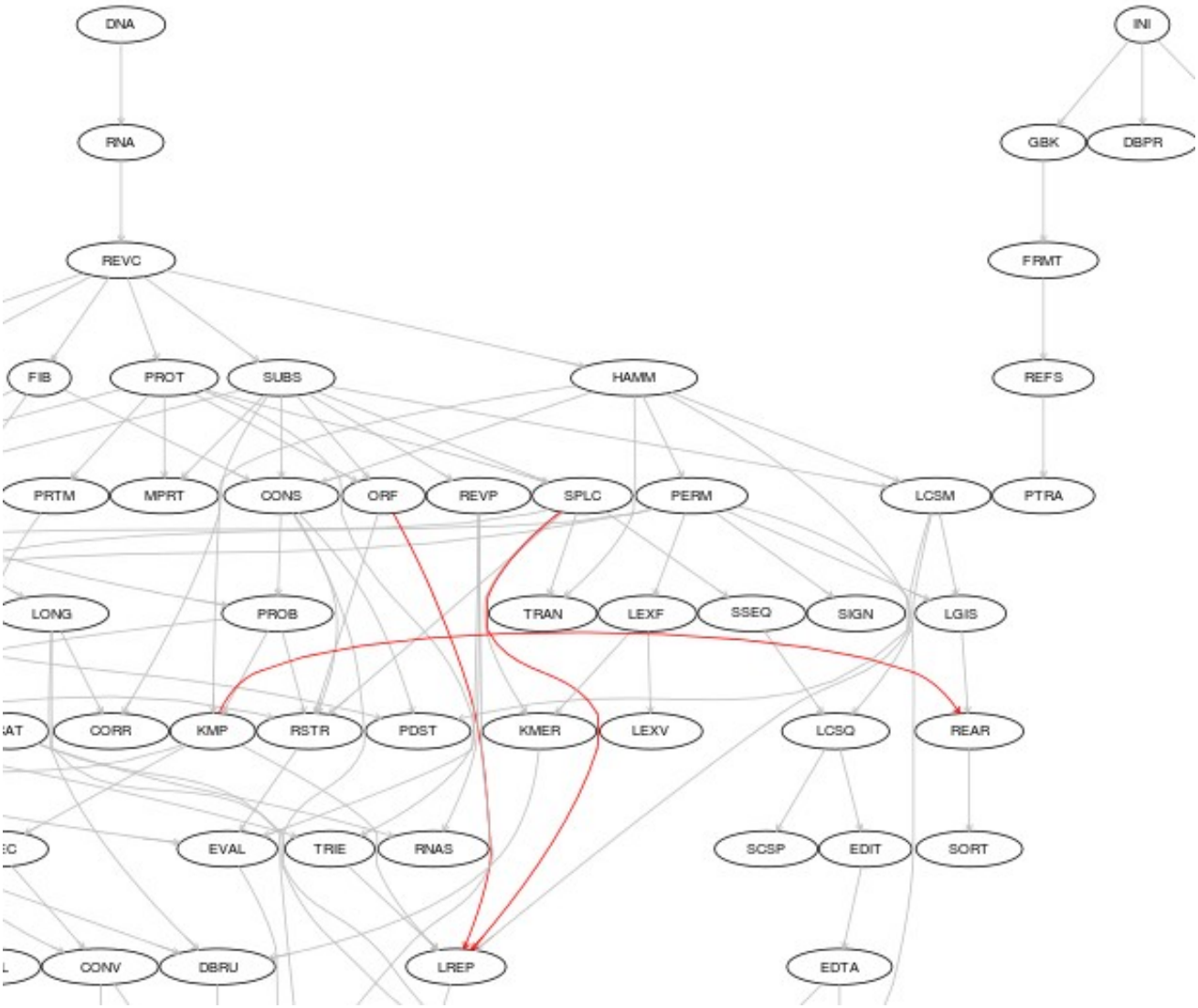

Рис. 2. Фрагмент графа с найденными дополнительными связями (обозначены красным: $KMP \to REAR, ORF \to LREP, SPLC \to LREP$).

Заключение

В ходе работы были выработан и реализован метод нахождения недостающих связей в графе задач с помощью анализа последовательностей решения задач отдельными пользователями, который даёт результаты, согласующиеся с обнаруженными пользователями проблемами.

Стоит заметить, что эти результаты носят чисто рекомендательный характер, и окончательное решение о внесении новых связей должно приниматься администраторами проекта. Для этого в дальнейшем планируется интегрировать реализованные алгоритмы в систему. Также в дальнейшие планы входит индивидуализация подхода, построение моделей студентов (или групп студентов, выделенных с помощью предварительной классификации), выделение успешных паттернов поведения, персонализация подсказок, плюс нахождение оптимальных путей для достижения поставленной обучающимся цели (например, пользователь может хотеть научиться решать конкретную задачу, находящуюся на сравнительно большой глубине в дереве, при этом его могут не интересовать сами промежуточные шаги, а лишь их количество, то есть затраченные усилия).

Ссылки и источники

[1] The State of Educational Data Mining in 2009: A Review and Future Visions. Ryan S.J.D. Baker, Kalina Yacef. Journal of Educational Data Mining, Volume 1, pp. 3-17. 2009.

[2] Mining Collaborative Patterns in Tutorial Dialogues. Sidney D'Mello, Andrew Olney and Natalie Person. Journal of Educational Data Mining, Volume 2, pp. 1-37. 2010.

[3] Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments. Saleema Amershi and Cristina Conati. Journal of Educational Data Mining, Volume 1, pp. 18-71. 2009.

[4] Identifying Successful Learners from Interaction Behaviour. Judi McCuaig, Julia Baldwin. 2012.

[5] Modeling Learning Patterns of Students with a Tutoring System Using Hidden Markov Models. Carole Beal, Sinjini Mitra and Paul R. Cohen. 2007.

[6] A Joint Probabilistic Classification Model of Relevant and Irrelevant Sentences in Mathematical Word Problems. Suleyman Cetintas, Luo Si, Yang Pin Xing, Dake Zhang, Joo Young Park and Ron Tzur. Journal of Educational Data Mining, Volume 2, pp. 83-101. 2010.

[7] Mining Diagnostic Assessment Data for Concept Similarity. Tara Madhyastha and Earl Hunt. Journal of Educational Data Mining, Volume 1, pp. 72-91. 2009.

[8] http://www.rosalind.info/

[9] http://www.coursera.org/

[10] http://www.khanacademy.org/about

[11] http://www.knewton.com/

[12] Knewton Adaptive Learning Whitepaper. http://www.knewton.com/wp-content/uploads/knewton-adaptive-learning-whitepaper.pdf