МЕТОДЫ ПОСТРОЕНИЯ И РАБОТЫ С КАТАЛОГОМ КНИЖНЫХ ИЗДАНИЙ

Юргель В.Ю. Email: Jurgel677@scientifictext.ru

Юргель Владислав Юрьевич - магистрант,
кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектроники,
г. Минск, Республика Беларусь

Аннотация: *в данной работе исследованы возможные подходящие модели обработки книжных изданий и общей текстовой информации для последующего поиска на основе пользовательских предпочтений и путем применения рассмотренных моделей для дальнейшего успешного составления соответствующего каталога с рекомендациями. Актуальность выбранной темы обусловлена стремительным ростом количества книжных изданий и, в общем, текста, что, тем самым, приводит к усложнению поиска рекомендаций в сети интернет и необходимости его максимального ускорения.*
Ключевые слова: *книги, рекомендации, нейронные сети, оптимизация, обработка текста.*

METHODS OF CONSTRUCTION AND WORK WITH THE CATALOG OF BOOK EDITIONS

Jurgel V.Yu.

Jurgel Vladislav Yurievich - Undergraduate,
INFORMATION TECHNOLOGY SOFTWARE DEPARTMENT,
BELARUSIAN STATE UNIVERSITY OF INFORMATICS AND RADIOELECTRONICS,
MINSK, REPUBLIC OF BELARUS

Abstract: *in this paper, possible suitable models of processing book editions and General text information for subsequent search based on user preferences and by applying their considered models for further successful compilation of the appropriate catalog with recommendations are investigated. The relevance of the chosen topic is due to the rapid growth of the number of books and, in General, the text, which, thereby, leads to the complexity of the search for recommendations on the Internet and the need to accelerate it as much as possible.*
Keywords: *books, recommendations, neural networks, optimization, text processing.*

УДК 004.021

Появление сети Интернет и бурный рост доступной текстовой информации значительно ускорило развитие научной области, существующей уже много десятков лет и известной как автоматическая обработка текстов (Natural Language Processing) и компьютерная лингвистика (Computational Linguistics). В рамках этой области предложено много перспективных идей по автоматической обработке текстов на естественном языке, которые были воплощены во многих прикладных системах, в том числе коммерческих. Сфера приложений компьютерной лингвистики постоянно расширяется, появляются все новые задачи, которые успешно решаются, в том числе с привлечением результатов смежных научных областей. Количество книжных изданий с каждым годом становится все больше, таким образом происходит довольно быстрый рост общего объема текстовой информации. Такая ситуация приводит к необходимости автоматизации формирования и поиска похожих книжных изданий на основе сформированных пользовательских предпочтений, что позволит моментально выполнять поиск книжных изданий и предлагать пользователю добавить в свой каталог для последующего приобретения. Одна из эффективных областей для выполнения автоматизации обработки текстов как раз и является компьютерная лингвистика и автоматическая обработка текстов.

Перед непосредственным выполнением интеллектуального поиска подходящих книжных изданий исходя из пользовательских предпочтений - пользователю необходимо наполнить личный каталог наиболее предпочитаемыми книжными изданиями.

Реализовать такой механизм довольно несложно, достаточно предоставлять пользователю кнопки «Нравится», и «Не нравится», дающие возможность системе понять, нравится книжное издание пользователю или нет по их нажатию соответственно. Однако не всегда достаточно просто определить степень эмоциональных переживаний при чтении книги, поэтому, помимо использования метода стандартных кнопок, дополнительно вводят систему рейтинга книжного издания и его комментирования. Таким образом мы можем сформировать полные вкусовые предпочтения пользователя по его каталогу и пользовательскому рейтингу отдельно взятой книги. Основной целью и задачей является исследование моделей и методов для автоматизации составления, категоризации и поиска похожих текстов в книжных изданиях для дальнейшего предоставления выбора пользователю на основе его заранее сформировавшихся предпочтений.

В наше время используется несколько методов обработки и построения каталога похожих книжных изданий:

- коллаборативная фильтрация - метод построения прогнозов (рекомендаций) в рекомендательных системах, который использует известные предпочтения (оценки) группы пользователей для прогнозирования неизвестных предпочтений другого пользователя;
- контентная фильтрация;
- корреляция Пирсона;
- алгоритмы кластеризации;
- классификация текста с помощью нейронных сетей.

Коллаборативная фильтрация.

Коллаборативная фильтрация вырабатывает рекомендации, основанные на модели предшествующего поведения пользователя. Эта модель может быть построена исключительно на основе поведения данного пользователя или, что более эффективно, с учетом поведения других пользователей со сходными характеристиками. В тех случаях, когда коллаборативная фильтрация принимает во внимание поведение других пользователей, она использует знание о группе для выработки рекомендаций на основе подобия пользователей. По существу рекомендации базируются на автоматическом сотрудничестве множества пользователей и на выделении (методом фильтрации) тех пользователей, которые демонстрируют схожие предпочтения или шаблоны поведения. В качестве примера предположим, что создается веб-сайт, чтобы предлагать его посетителям рекомендации относительно различных тематик (блогов, книжных жанров и т.д.). На основе информации от многих пользователей, которые подписываются на различные тематики по книгам и читают их, можно сгруппировать этих пользователей по их предпочтениям. Например, можно объединить в одну группу пользователей, которые читают несколько одних и тех же книг. По этой информации происходит идентифицирование самых популярных жанров книг среди тех, которые читают участники этой группы. Затем, конкретному пользователю этой группы, необходимо рекомендовать самый популярный книжный жанр из тех, на которые он еще не подписан и которые он не читает.

Контентная фильтрация.

Контентная фильтрация формирует рекомендацию на основе поведения пользователя. Например, этот подход может использовать ретроспективную информацию о просмотрах (какие блоги читает пользователь и характеристики этих блогов). Если какой-либо пользователь обычно читает статьи о Linux или регулярно оставляет комментарии в блогах по проектированию программного обеспечения, то контентная фильтрация может использовать эту ретроспективную информацию для выявления подобного контента и предложения такого контента в качестве рекомендованного для этого пользователя (статьи в блогах по Linux или в других блогах по проектированию программного обеспечения). Этот контент может быть определен в ручном режиме или извлечен автоматически на основе других методов подобия [1, с. 12].

Корреляция пирсона.

Сходство между двумя пользователями и их атрибутами, такими как статьи, прочитанные в коллекции блогов, может быть точно вычислено с помощью так называемой корреляции Пирсона. Этот алгоритм измеряет линейную зависимость между двумя переменными или пользователями как функцию их атрибутов. Однако он не вычисляет эту меру по всей совокупности пользователей. Эту совокупность необходимо предварительно отфильтровать до близких элементов на основе высокоуровневых показателей сходства, таких как чтение похожих тематик или жанров. Корреляция Пирсона, которая широко применяется в исследовательской деятельности, является весьма популярным алгоритмом в сфере коллаборативной фильтрации.

Алгоритмы кластеризации.

Алгоритмы кластеризации - это разновидность "спонтанного обучения", позволяющая выявить структуру в рядах на первый взгляд случайных данных. В общем случае такой алгоритм базируется на выявлении сходства между элементами (например, между читателями блога) посредством вычисления их расстояния от других элементов в пространстве признаков (признаком в пространстве признаков может, например, быть количество прочитанных статей в наборе блогов). Количество независимых признаков определяет размерность пространства признаков. Если элементы "близки" друг к другу, то их можно объединить в один кластер. Существует множество алгоритмов кластеризации. Самым простым из них является алгоритм k-средних, который разделяет элементы на k кластеров. Первоначально элементы распределяются по этим кластерам в произвольном порядке. Затем для каждого кластера вычисляется центр масс (или просто центр) как функция его членов. После этого проверяется расстояние каждого члена кластера от центра этого кластера.

Список литературы / References

1. *Большакова Е.И.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 269 с.
2. *Васильев В.Г., Кривенко М.П.* Методы автоматизированной обработки текстов. М.: ИПИ РАН, 2008. 301 с.
3. *Гладкий А.В.* Синтаксические структуры естественного языка в автоматизированных системах общения. М.: Наука, 1985. 144 с.