УДК 004.852

Ускорение работы рекомендательной системы с применением методов обработки естественных языков

Д.И. Борисяк
Московский физико-технический институт (государственный университет)

Рассмотрен алгоритм рекомендаций литературы пользователям электронной библиотеки, основанный на нахождении скрытых предпочтений пользователя подобно методам, основанным на неотрицательной матричной факторизации, с фиксированием матрицы объектов (литературных произведений). Матрицы объектов получены при помощи методов Word2Vec[1] и Doc2Vec[2], которые ставят в соотвествие словам/тексту их векторное представление. Векторы Doc2Vec были получены непосредственно из содержания литературных произведений, в то время как в модели Word2Vec рассматривались последовательности прочтения книг каждым пользователем. В большинстве практических задач похожим текстам ставятся в соответствие близкие векторы, что делает возможным построение рекомендаций по этим векторам.

Стоит отметить, что так как векторы объектов в традиционных методах, основанных на матричных разложениях, оцениваются из данных (скрытые переменные), их замена на фиксировнные векторы Doc2Vec (получаемые напрямую из содержания произведения) позволяет улучшить качество оценки скрытых параметров пользователя за счет уменьшения общего числа оцениваемых параметров при неизменном размере обучающей выборки.

Другим приемуществом рассматриваемого метода является значительное ускорение процесса обучения за счет уменьшения размерности задачи. Особенно это является актуальным при работе с большими данными, который является частым для рекомендательных систем в индустрии. Методы наподобие Alternating Least Squares (поиск скрытых параметров как объектов, так и пользователей) требуют передачи данных между вычислительными узлами после каждой раунда оптимизации, в то время как при фиксировании матрицы объектов задача разбивается на множество независимых подзадач, что снижает к минимуму затраты на обмен данными между вычислительными узлами.

Обозначим векторы, полученные алгоритмами Word2Vec и Doc2Vec, как B, матрица размера $N \times k$. Предположим, что скрытые параметры i-го пользователя w_i - вектор длины k, а $j-$ая компонента вектора y_i соответствует прочтению j-ой книги пользователем: 1 при прочтении, 0 иначе. В отличие от классических алгоритмов основанных на матричных разложениям, здесь рассматривается задача классификации (прочитана/не прочитана), а ошибки классификации (книга не прочитана при предсказании, что она должна быть прочитана) расцениваются как потенциальные рекомендации. В качестве классификатора рассматривается логистическая регрессия - модель оценки апостериорной вероятности класса:

$$p(y_i^j | x) = expit(w_i B_i + b_i)$$

где $expit(\cdot)$ - логистическая функция.

Логистическая регрессия минимизирует следующую функцию потерь:

$$L_i = \sum_{j=1}^{N} y_i^j log\, p(y_i^j | x) + (1 - y_i^j)\, log\,(1 - p(y_i^j | x)) + C \left| w_i \right|_2^2$$

где $\left| w_i \right|_2^2$ - квадрат l_2-нормы вектора w_i.

Обучая логистическую регрессию для каждого пользователя, метод определяет скрытые параметры пользователя и способ предсказания вероятности прочтения книги.

Таким образом, предложенный метод позволяет ускорить и повысить качество рекомендательной системы для электронной библиотеки за счет учета информации о содержании самих книг.

Оценка качества классификатора приведена на рис. 1. На рис. 2 представлено распределение полученных значений скрытых переменных в векторах пользователей. Для обработки данных

использовался язык программирования Python с применением библиотек Theano, Scikit-learn и Lasagne.

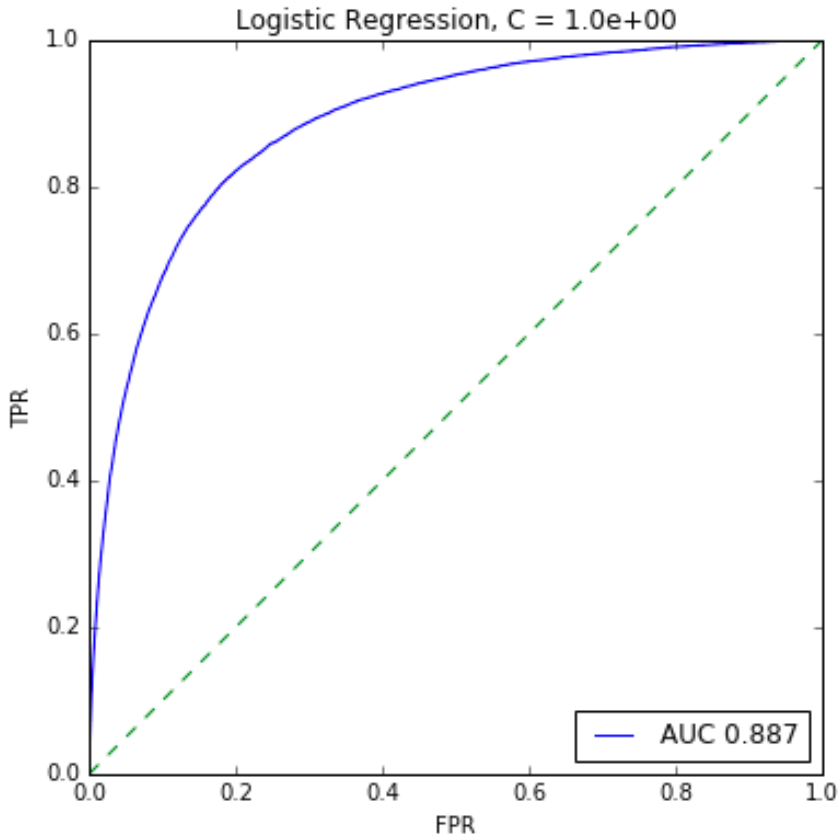

Рис.1 Оценка качества классификации при помощи ROC-кривой

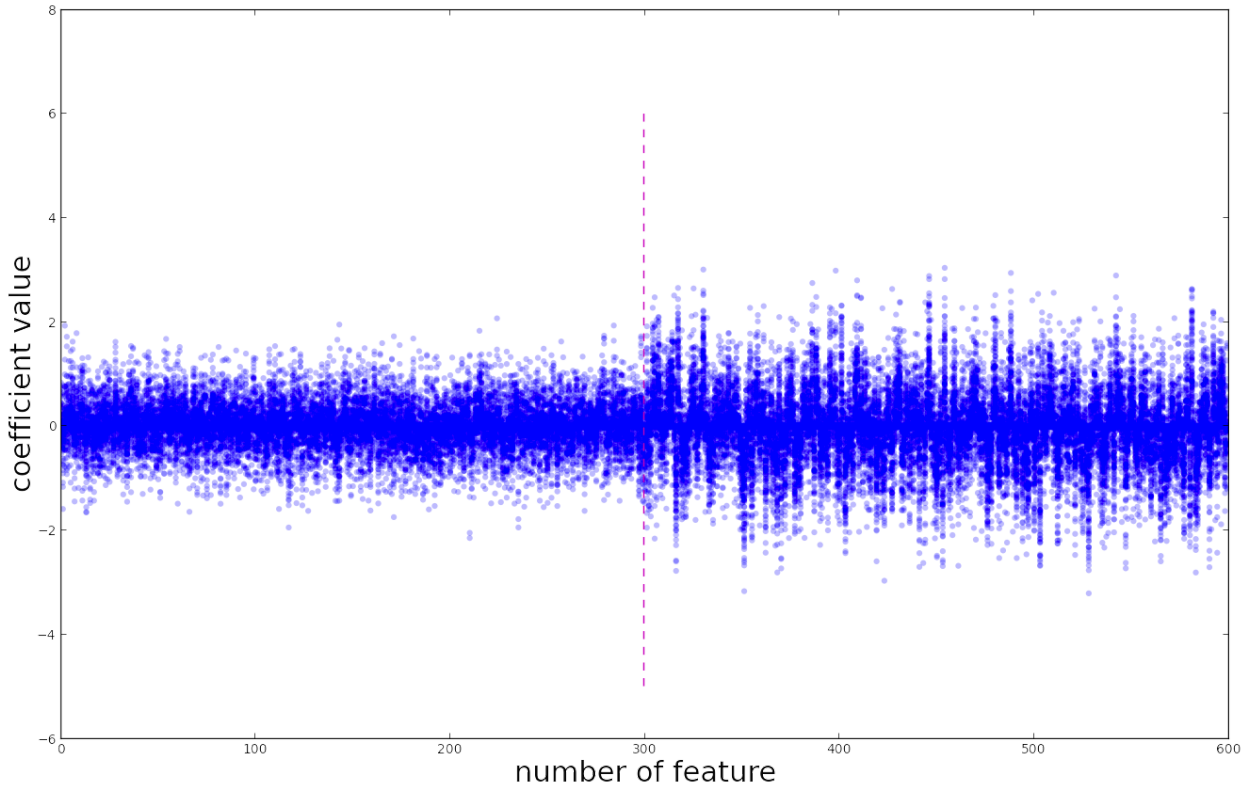

Рис.2 Распределение полученных значений скрытых переменных в векторах пользователей

Литература

1. *Mikolov T.* et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. – 2013. – С. 3111-3119.
2. Distributed Representations of Sentences and Documents //ICML. – 2014. – Т. 14. – С. 1188-1196.
3. *Zhou Y.* et al. Large-scale parallel collaborative filtering for the netflix prize //International Conference on Algorithmic Applications in Management. – Springer Berlin Heidelberg, 2008. – С. 337-348.