Рекомендация на основе мультимодальных данных из нескольких источников

Иван Самборский
Университет ИТМО
Санкт-Петербург, Россия
samborsky@rain.ifmo.ru

Денис Механиков
Университет ИТМО
Санкт-Петербург, Россия
mehanikov@rain.ifmo.ru

Александр Фарсеев
Национальный Университет Сингапура
Сингапур
farseev@u.nus.edu

Аннотация—В последние годы, социальные сети играют решающую роль в повседневной жизни индивидуумов. С каждым днем увеличивается число пользователей, использующих несколько социальных сетей одновременно. Исходя из этого, стоит предположить что данные из нескольких социальных сетей объединенные в одной модели могут повысить производительность систем, решающих такие задачи как профилирование пользователей и рекомендации. Для поддержки вышеуказанных суждений, в этой работе мы описываем рекомендательную систему, использующую несколько социальных сетей одновременно, и доказываем, что объединение данных из различных источников позволяет достичь более высокой производительности при решении задачи рекомендации по сравнению с существующими подходами.

I. Введение

За последние несколько лет произошел огромный скачек объемов общедоступной информации в Интернете, что привело к существенным проблемам в сфере извлечении полезных знаний. Одним из решений данной проблемы являются рекомендательные системы, которые способны фильтровать не релевантную информацию. Кроме того, многие пользователи одновременно используют две и более социальных сети ежедневно [3], что позволяет предположить, что интересы таких пользователей могут быть определены лучше с использованием данных из различных источников, описывающих пользователей с различных ракурсов. В то же время, ранее было установлено, что эффективная рекомендация возможна при комбинации групповых и индивидуальных знаний [2], поскольку индивидуальные знания описывают личный опыт пользователя [9], в то время как, групповые — помогают повысить разнообразие рекомендуемой выборки [7].

В рамках данной работы, мы описываем систему предназначенную для рекомендации категории места (например: магазин одежды, отель), которая объединяет групповые и индивидуальные знания для осуществления рекомендации основанной на данных из различных социальных сетей. Индивидуальные знания представлены в виде прошлого опыты пользователя — распределения по категориям мест, которые уже были им посещены. Групповые знания извлекаются из сообщества пользователя, определенного автоматически на основе мультимодальных данных. Рекомендуемые категории мест были взяты из социальной сети Foursquare, которых, на момент сбора данных, было 764.

II. Корпус данных

Исследование проводилось на основе данных из корпуса NUS-MSS [6]. Корпус содержит данные из трех социальных сетей: Foursquare, Instagram, и Twitter —, собранных в трех городах: Сингапур, Лондон, и Нью-Йорк.

III. Подход к рекомендации

Принятие решений, как известно, основывается на двух факторах: личный опыт и общественное мнение [1]. Общественное мнение может быть определено сообществом пользователя, к которому он принадлежит. Это явление может быть использовано для повышения эффективности рекомендации. Исходя из этого мы производим рекомендацию на основе как личных, так и групповых знаний, которые естественным образом моделируют влияние общества на поведение индивида при выборе нового места. Формально подход к рекомендации может быть описан следующим образом:

$$rec(u) = sort(\gamma \cdot vec_u + \theta \frac{\sum_{v \in C_u} vec_v}{|C_u|}),$$

где vec_u распределение пользователя u по категориям мест, а правая часть — распределение всех членов сообщества по категориям мест.

A. Выделение сообществ пользователей на основе одного источника данных

Одной из наиболее распространенных формулировок задачи определения сообществ является ее представление в виде проблемы MinCut [10], которая для заданного числа k, заключается в нахождении подмножеств $C_1, ..., C_k$ таких, что они минимизируют выражение $cut(C_1, ..., C_k) = \sum_{i=1}^{k} W(C_i, \overline{C}_i)$, где W — функция расстояния. В нашем случае — сумма весов всех ребер подграфа.

MinCut проблема является $\mathcal{NP}$-трудной [11], но может быть приближенно решена с помощью, так называемой, спектральной кластеризации (spectral clustering) [10]: $\min_{U \in R^{n \times k}} \mathrm{tr}(U^\intercal L_{sym} U)$, где $U^\intercal U = I$.

B. Определение сообществ пользователей на основе нескольких источников данных

Стремясь повысить производительность рекомендации, мы разработали подход для определения сооб-

ISBN 978-952-68397-6-9 (paperback), ISBN 978-952-68397-7-6 (PDF)

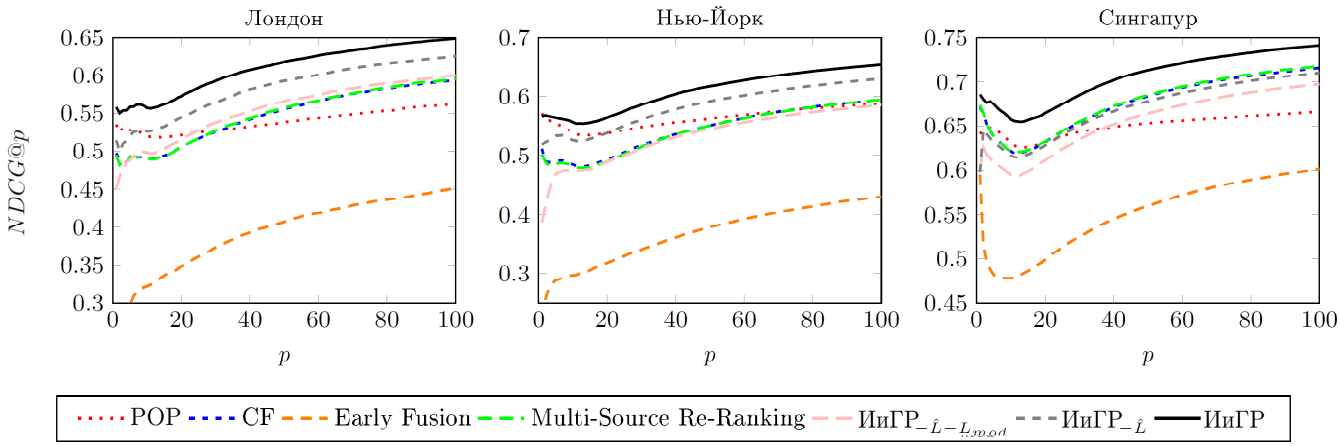

Рис. 1: Эффективность рекомендации нашей системы в сравнение с другими алгоритмами

ществ пользователей на основе данных из нескольких источников. Для представления связей между пользователями мы используем многослойный граф [4], каждый слой которого построен на основе косинусного расстояния между векторами признаков пользователей из этого источника данных. Таким образом, целевая функция принимает следующий вид:

$$\min_{U \in \mathbb{R}^{n \times k}} \operatorname{tr}(U^\intercal \sum_{i=1}^{M} (\hat{L}_i - \alpha \hat{U}_i \hat{U}_i^\intercal) U),$$

где $\hat{L}_i := L_i - \beta_i \sum_{j=1, j \neq i}^{M} w_{i,j} U_j U_j^\intercal$, а $w_{i,j}$ — схожесть между источниками i и j.

IV. Результаты

Для сравнения соперничающих рекомендательных систем между собой, корпус данных был поделен на две части: тренировочную выборку — первые 3 месяца, и тестовую выборку — оставшиеся 2 месяца. В выборки были включены только те пользователи, для которых есть данные из всех трех социальных сетей. В итоге мы получили: 1801 пользователя из Сингапура, 813 из Лондона и 1602 из Нью-Йорка.

В качестве оценки использовалась мера $NDCG@p$ (Normalized Discounted Cumulative Gain), которая определена как $NDCG@p = \frac{DCG@p}{IDCG@p}$, где $DCG@p = \sum_{i=1}^{p} \frac{2^{rel_i}}{\log_2(i+1)}$, $IDCG$ — максимально возможное значение (Ideal) DCG для конкретного запроса, rel_i — релевантность результата на позиции i ($rel_i = \frac{Cat_j}{N_{Cat}}$, где Cat_j — количество посещений мест, принадлежащих к категории i, $N_{Cat} = 764$ — общее количество категорий Foursquare).

В качестве алгоритмов-соперников были взяты следующие подходы к рекомендации:

- Popular (POP) — производит рекомендацию основываясь только на прошлом опыте пользователя;
- Collaborative Filtering (CF) [8] — рекомендует на основании поведения k наиболее похожих пользователей из обучающей выборки;
- Early Fusion [12] — объединяет данные нескольких источников в единый вектор признаков; далее на полученных данных используется CF;
- Multi-Source Re-Ranking [5] — линейная комбинация результатов нескольких источников, с использованием заранее подобранных весов.

А также частные случаи нашего подхода — ИиГР (Индивидуальная и Групповая Рекомендация):

- $_{-\hat{L}}$ — при $\beta_i = 0, i = 1..M$;
- $_{-\hat{L}-L_{mod}}$ — при $\beta_i = 0, i = 1..M$ и $\alpha = 0$.

Полученные результаты сравнения эффективности рекомендации относительно базовых алгоритмов представлены на Рис. 1. Из рисунка видно, что слияние нескольких источников данных в единую рекомендательную систему может значительно повысить производительность рекомендации по сравнению с другими подходами.

V. Заключение

В данной работе мы исследовали влияние мультимодальных данных из различных источников социальных медиа для решения проблемы рекомендации. На основании корпуса данных NUS-MSS [6], мы показали что слияние нескольких источников данных в единую рекомендательную систему может значительно повысить производительность рекомендации категорий мест.

Наши будущие исследования включают в себя разработку полуобучаемой модели для определения сообществ пользователей на основе знаний о предметной области. Кроме того, мы планируем использовать дополнительные источники данных, такие как социальный граф и данные с носимых устройств, в целях дальнейшего повышения производительности рекомендации.

Благодарность

Эта работа поддержана правительством Российской Федерации, грант 074-U01 и Singapore National Research Foundation в рамках International Research Centre @ Singapore Funding Initiative, управляемым IDM Programme Office.

Список литературы

[1] A. Ambrus, B. Greiner, P. Pathak, et al, “Group versus individual decision-making: Is there a shift”.

[2] R. Burke, “Integrating knowledge-based and collaborative-ltering recommender systems”, Workshop on AI and Electronic Commerce, 1999.

[3] P. R. Center, “Social networking fact sheet 2014”, Web: www.pewinternet.org/fact-sheets/social-networking-fact-sheet.

[4] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, “Clustering on multi-layer graphs via subspace analysis on grassmann manifolds”, Signal Processing, IEEE Transactions, 2014.

[5] A. Farseev, D. Kotkov, A. Semenov, J. Veijalainen, and T.-S. Chua, “Cross-social network collaborative recommendation”, ACM International Conference on Web Science, 2015.

[6] A. Farseev, L. Nie, M. Akbari, and T.-S. Chua, “Harvesting multiple sources for user profile learning: a big data study”, 5th ACM on International Conference on Multimedia Retrieval, 2015.

[7] R. Hu and P. Pu, “Helping users perceive recommendation diversity”, Workshop on novelty and diversity in recommender systems, 2011.

[8] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems handbook”, ESpringer, 2011.

[9] S. Trewin, “Knowledge-based recommender systems”, Encyclopedia of Library and Information Science, vol.69, 2000.

[10] U. Von Luxburg, “A tutorial on spectral clustering. Statistics and computing”, 2007.

[11] D. Wagner and F. Wagner, “Between min cut and graph bisection”, Springer, 1993.

[12] P. Winoto and T. Tang, “If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? a study of cross-domain recommendations”, New Generation Computing, 2008.