

УДК 004.942

МЕТРИКИ, НЕОБХОДИМЫЕ ДЛЯ СОСТАВЛЕНИЯ РАНЖИРОВАННОГО СПИСКА НОВОСТЕЙ ДЛЯ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ

Колебцев В.И., Белов Ю.С.

*Московский государственный технический университет имени Н.Э. Баумана, филиал,
Калуга, e-mail: fn1-kf@mail.ru*

Проблема фильтрации и рекомендации новостей исследуется уже долгие годы. Технически в эти годы было выдвинуто множество подходов, от подходов коллаборативной фильтрации до методов, основанных на содержании, или гибридных систем. Главная цель персонализированной рекомендации новостей – помочь читателям найти интересные истории, максимально соответствующие их читательским интересам. Однако новостная область обладает рядом характеристик, которые делают задачу рекомендации особенно трудной: ускоренное снижение релевантности статьи (релевантность статьи может очень быстро уменьшиться после публикации или при появлении новой информации по текущей тематике), быстро растущее количество статей (ежедневно на новостных порталах добавляются сотни новых новостных статей. Это усиливает проблему холодного запуска. Тем не менее свежие элементы должны рассматриваться для рекомендации, даже если для них записано не слишком много взаимодействий), смещение предпочтений пользователей (предпочтения отдельных пользователей часто не так стабильны, как в других областях, например, развлечения). В данной статье описаны основные факторы, формирующие контекст статьи. Учитывая эти факторы, можно максимально оптимизировать вычисление актуальности статьи. Также в статье представлен возможный способ балансировки точности рекомендации с учетом новизны статьи.

Ключевые слова: рекомендация новостей, метрики новостей, точность рекомендации, актуальность элемента рекомендации

METRICS REQUIRED TO COMPILE A RANKED LIST OF NEWS FOR A RECOMMENDATION SYSTEM

Kolebtsev V.I., Belov Yu.S.

Bauman Moscow State Technical University, Kaluga branch, Kaluga, e-mail: fn1-kf@mail.ru

The problem of filtering and recommending news has been studied for many years. Technically, many approaches have been put forward in these years, from collaborative filtering approaches to content-based methods or hybrid systems. The main goal of a personalized news recommendation is to help readers find interesting stories that best match their reader interests. However, the news area has several characteristics that make the task of recommendation particularly difficult: an accelerated decline in the relevance of the article (the relevance of the article can decrease very quickly after publication or when new information on the current topic appears), a rapidly growing number of articles (hundreds of new news articles are added daily on news portals. This reinforces the cold start problem. However, fresh items should be considered for recommendation, even if not too many interactions are recorded for them), user preference bias (individual user preferences are often not as stable as in other areas, such as entertainment). This article describes the main factors that form the context of the article. Given these factors, it is possible to optimize the calculation of the relevance of the article as much as possible. The article also presents a possible way to balance the accuracy of the recommendation with the novelty of the article.

Keywords: news recommendation, news metrics, recommendation accuracy, recommendation element relevance

Благодаря технологии персонализированных рекомендаций, особенно популярных сегодня в области электронной коммерции, они также постепенно распространились на другие приложения, в том числе и персонализированные рекомендации новостей, которые обсуждаются в данной работе. Однако, поскольку новости имеют свои уникальные характеристики, ключевые технологии персонализированной рекомендации новостей немного отличаются от таковых в традиционной сфере электронной коммерции [1]. Ниже приведено сравнение и анализ ключевых технологий в персонализированной рекомендации новостей с разных точек зрения характеристик новостей.

Цель исследования: изучить факторы, влияющие на актуальность статьи, и при-

вести структуру зависимости этих параметров друг от друга. На основе полученной модели описать исчерпывающую информацию об основных факторах, от которых зависит актуальность новостной статьи, и предложить способ балансировки точности рекомендации статьи с учетом параметра новизны.

Основные факторы, влияющие на актуальность статьи. Новости используются для быстрого и своевременного сообщения людям последних, своевременных и, главное, «ценных» фактов с помощью кратких текстов. «Недавний» фактор относится к контекстуальной информации новостей в смысле времени (т.е. когда новость произошла), а «ближайший» фактор относится к контекстуальной информации новостей в смысле местоположения (т.е. где новость

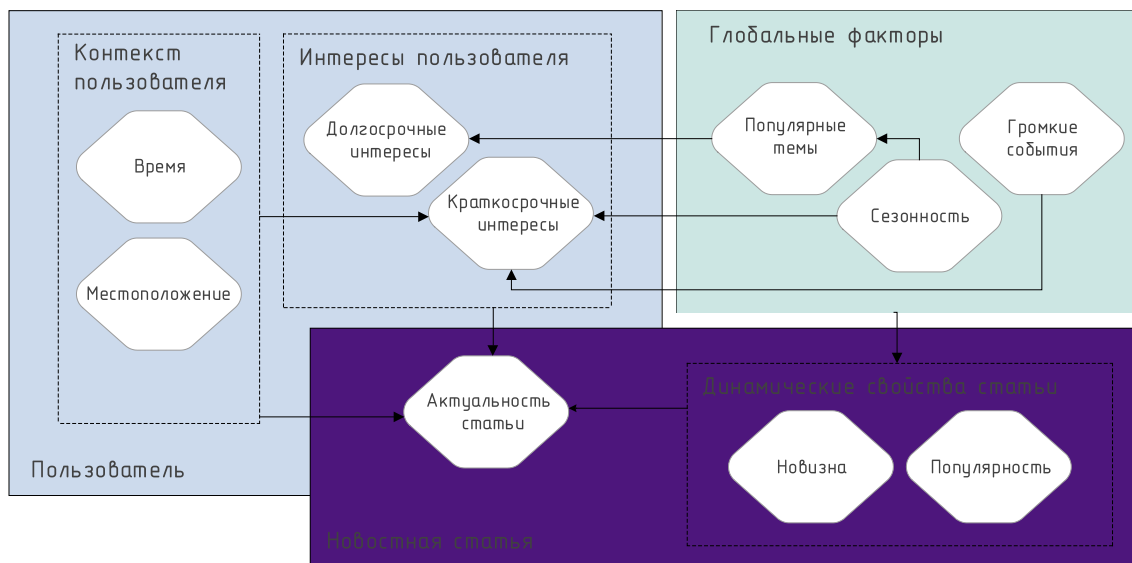
произошла) [2]. Эти два фактора являются основной контекстуальной информацией, рассматриваемой в персонализированной новостной рекомендации. Кроме того, существует множество других контекстуальных факторов, таких как погода, сезон, будний день или выходные, плотность толпы и наличие сверстников среди окружения. В этой работе обсуждается только влияние контекстной информации о времени и местоположении на персонализированные рекомендации новостей. С популяризацией портативных мобильных устройств, таких как мобильный телефон, iPad и ноутбук, читатели новостей постепенно переходят с веб-приложений на мобильные. Благодаря портативности и гибкости мобильных устройств пользователи могут читать новости в любом месте и в любое время. Поэтому существует острая необходимость в изучении технологии персонализированных рекомендаций новостей, которая интегрирует контекстуальную информацию о времени и местоположении.

Учитывая все вышеописанное, на рисунке показана концептуальная модель факторов релевантности новостей. В этой модели на релевантность новостной статьи для отдельного пользователя может влиять ряд факторов, в том числе факторы, связанные со статьей, связанные с пользователем и так называемые глобальные факторы. Ниже подробно рассмотрены основные факторы.

Цель персонализированной рекомендации новостей, зависящей от времени и ме-

стоположения, состоит в том, чтобы точно предсказать интересы пользователей к чтению в определенный момент времени (или период) и в определенном месте [3]. Если взять в качестве примера задачу рекомендации $Top - N$, то цель состоит в том, чтобы в итоге сформировать список новостей длиной N для пользователей, с учетом того, что список должен содержать только новости, интересные пользователю в определенный момент или в определенном месте. Детальный анализ технологии рекомендаций новостей с учетом фактора времени и фактора местоположения выглядит следующим образом.

Фактор времени. Влияние временной контекстуальной информации на персонализированную систему рекомендаций новостей и интересы пользователей является обширным и глубоким и в основном отражается в трех аспектах. Во-первых, интересы пользователей постоянно меняются. В зависимости от конкретного пользователя его интересы могут меняться постепенно или резко с течением времени. Для того чтобы точно извлечь текущие процентные предпочтения пользователей с постепенными изменениями интереса, их недавнее поведение должно быть также рассмотрено, поскольку недавнее поведение может наилучшим образом отражать текущие процентные предпочтения [4]. Во-вторых, новости имеют свой жизненный цикл. Жизненный цикл новостей короткий. Сегодняшние «горячие» новости могут потерять свою актуальность завтра.



Концептуальная модель факторов релевантности новостной статьи

Поэтому при рассмотрении вопроса о рекомендации новостей пользователям в данный момент необходимо различать, являются ли новости устаревшими или нет. В-третьих, новости имеют эффект сезонности. Например, различные громкие события, повторяющиеся с какой-то периодичностью, будут влиять на интересы пользователей, что является эффектом, вызванным изменением сезона. В общем случае при разработке алгоритмов рекомендаций необходимо учитывать данный фактор.

В связи с тем, что интересы пользователей меняются с течением времени, будет целесообразно делить интересы пользователей на два типа – долгосрочные предпочтения и краткосрочные предпочтения. Долгосрочные предпочтения пользователя относятся к общим предпочтениям, которые пользователь сформировал от начала использования системы до текущего времени. Краткосрочные предпочтения в основном основаны на недавнем поведении пользователя при чтении. Затем различные группы новостей, предпочитаемые разными пользователями, могут быть разделены, учитывая долгосрочные предпочтения пользователей, а затем с помощью краткосрочных предпочтений этот метод позволяет выбрать именно те новости, которые, скорее всего, будут актуальны для пользователя в данный момент времени. Этот метод полностью сочетает долгосрочные интересы пользователя с краткосрочными интересами, полностью учитывая характеристику интересов пользователей, изменяющихся постепенно с течением времени. Еще один важный фактор – продолжительность времени чтения. То есть это время, которое пользователи тратят на чтение новостей. Используя его, система может различать пользователей, которые внимательно читают новости, и пользователей, которые читают новости поверхностно. Корректировка традиционной формулы оценки пользователей с помощью данного фактора может в определенной степени повысить точность рекомендаций. Однако этот метод не учитывает внешние факторы, влияющие на чтение. Например, продолжительность чтения может увеличиться, если пользователь прервался и продолжил чтение через определенный промежуток времени.

В персонализированной рекомендации новостей временной контекст играет очень важную роль в методе рекомендации пользовательской коллаборативной фильтрации [5], особенно для измерения сходства. После нахождения похожей группы пользователей для пользователя u , краткосрочные предпочтения пользователей в этой группе, по-видимому, ближе к текущим новостным

темам пользователя u , чем новостные темы, интересующие их некоторое время назад. Иными словами, на последние прочитанные новости пользователей, имеющих схожие интересы с целевым пользователем u , должно быть обращено больше внимания при рекомендации новостей целевому пользователю. Это особенно важно в области новостных рекомендаций с очень короткими жизненными циклами. Учитывая все вышеописанное, ниже приведена формула расчета сходства пользователей (модернизированная формула в алгоритме user-based коллаборативной фильтрации) (1).

В уравнении (1) к числителю формулы добавляется коэффициент затухания времени λ для новостей, которыми поделились пользователь u и пользователь v . $N(u)$ и $N(v)$ означает коллекцию новостей, которую читают пользователь u или v соответственно. t_{ui} представляет собой временную точку отсчета i для пользователя u . Чем дальше временные точки i для u и v при чтении новостей, тем меньше их сходство. В результате найденные похожие друг на друга группы уникальных пользователей будут более точными.

$$sim_{uv} = \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{1 + \lambda |t_{ui} - t_{vi}|}}{\sqrt{|N(u)| |N(v)|}}. \quad (1)$$

Фактор местоположения. Персонализированная новостная рекомендация, рассматривающая информацию о местоположении, в основном рассматривает траекторию движения пользователя, а также построение и распознавание сети часто посещаемых мест. В частности, точки местоположения пользователя, изменяющиеся с течением времени, соединяются последовательно для формирования траектории движения. Это повторяется для каждого пользователя, зарегистрированного в системе. После этого анализируются различные маршруты относительно позиции всех пользователей на основе сети траекторий, чтобы сформулировать новый метод расчета ближайшего пользователя [6]. Даже при рассмотрении информации о перемещении только одного пользователя, будущие предпочтения в тематике новостей можно предсказать в соответствии с собственными траекториями перемещения конкретного пользователя. Затем пользователю будут предложены новости, которые, скорее всего, будут ему интересны в диапазоне прогнозируемой позиции, чтобы реализовать основанную на местоположении персонализированную рекомендацию новостей.

Фактор оценки новизны новостной статьи. Новизна рекомендуемого элемента может быть определена по-разному, например как неочевидность предложений элемента, или с точки зрения того, насколько отличается элемент по отношению к тому, что уже было предложено и оценено пользователем или группой пользователей. Однако рекомендация исключительно новых или непопулярных товаров может иметь ограниченную ценность, если они не соответствуют интересам пользователей [7]. Поэтому цель рекомендательной системы часто состоит в том, чтобы сбалансировать эти конкурирующие факторы, то есть сделать несколько более новые и, следовательно, рискованные рекомендации, в то же время обеспечивая высокую точность.

В литературе был предложен ряд способов количественной оценки степени новизны, в том числе альтернативные способы учета информации о популярности или расстоянии элемента-кандидата до профиля пользователя. Например, можно измерять новизну как противоположность популярности статьи, исходя из предположения, что менее популярные статьи с большей вероятностью будут неизвестны пользователям и их рекомендации приведут к более высоким уровням новизны.

Что касается рассмотрения компромиссных ситуаций, то возможны различные технические подходы. Можно, например, попытаться повторно ранжировать оптимизированный по точности список рекомендаций либо для достижения глобально определенных уровней качества, либо для получения списков рекомендаций, соответствующих предпочтениям отдельных пользователей. Другой подход заключается в варьировании весов различных факторов, чтобы найти конфигурацию, которая приводит как к высокой точности, так и к хорошей новизне.

Наконец, можно попытаться включить рассмотрение компромиссов в фазу обучения, например, используя соответствующий термин регуляризации. Существует метод под названием Novelty-aware Matrix Factorization (NMF), который пытается одновременно рекомендовать точные и новые элементы. Предлагаемый ими подход

к регуляризации носит точечный характер, что означает, что новизна каждого элемента-кандидата рассматривается индивидуально.

Балансировка точности рекомендации с учетом новизны. Конечной задачей системы рекомендации новостей является составление ранжированного списка элементов (Top-N рекомендация), которые, по предсказанию системы, пользователь будет читать следующими [8]. В связи с этим определим функцию потерь на основе ранжирования для поставленной задачи.

Введем термин регуляризации новизны в функцию потерь с целью включения аспекта новизны статьи непосредственно в процесс обучения модели. Этот параметр регуляризации может быть настроен для достижения баланса между точностью рекомендации и новизной предлагаемой статьи в соответствии с желаемым эффектом для конкретного приложения.

В данной статье предлагаемый подход определения новизны основан на методе непопулярности элемента. То есть, как писалось выше, можно измерять новизну как противоположность популярности статьи, исходя из предположения, что менее популярные статьи с большей вероятностью будут неизвестны пользователям и их рекомендации приведут к более высоким уровням новизны [9].

Таким образом, предлагаемый фактор новизны направлен на смещение рекомендации нейронной сети в сторону новых элементов в системе. В таком случае это не повлияет негативным образом на элементы из положительной выборки (те, на которые пользователи кликнули), а только на элементы из отрицательной (которые еще не были просмотрены). Новизна отрицательных элементов оценивается по их вероятности быть следующей предложенной статьей в последовательности (апостериорная вероятность того, что статья будет прочитана следующей во время текущего пользовательского сеанса), чтобы выдвинуть эти элементы на первое место в списке рекомендованных статей, которые будут одновременно новыми и релевантными для конкретного пользователя.

Функция потерь для фактора новизны определяется в уравнении

$$nov_loss(\theta) = \frac{1}{|C|} \sum_{(s, i^+, D^+) \in C} \frac{\sum_{i \in D^+} P(i | s, D^+) \cdot novelty(i)}{\sum_{i \in D^+} P(i | s, D^+)}, \quad (2)$$

где C – множество кликов пользователя, доступных для обучения, D^+ – случайные элементы из отрицательной выборки,

за исключением положительной выборки, i – идентификатор статьи, s – текущий пользовательский сеанс. Значение новизны

элементов оценивается по их прогнозируемой релевантности $P(i|s, D^t)$, чтобы поднять к вершине списка как новые, так и релевантные элементы.

$$\text{novelty}(i) = -\log_2(\text{rec_norm_pop}(i) + 1). \quad (3)$$

Метрика новизны в уравнении (3) определяется на основе нормализованной популярности статьи за последнее время. Отрицательный логарифм в уравнении увеличивает значение метрики новизны для long-tail элементов (элементы, которые мало кто видел).

$$\text{rec_norm_pop}(i) = \frac{\text{recent_clicks}(i)}{\sum_{j \in I} \text{recent_clicks}(j)}. \quad (4)$$

Вычисление нормализованной популярности статьи суммируется до максимального значения 1.0 для всех рекомендуемых элементов (набор I), как показано в уравнении (4). Поскольку балансировка актуальна только для тех статей, которые стали популярны недавно, в расчетах рассматриваются только клики, полученные статьей в течение определенного периода времени (например, за последний час). Данная информация возвращается функцией $\text{recent_clicks}(\cdot)$.

Результаты исследования и их обсуждение

В процессе написания данной статьи были изучены факторы, влияющие на актуальность статьи, и приведена структура зависимости этих параметров друг от друга. На основе полученной модели описана исчерпывающая информация об основных факторах (таких как время и связанные с этим профили предпочтений (долгосрочный и краткосрочный), местоположение и новизна новостной статьи), от которых зависит актуальность новостной статьи, и предложен способ балансировки точности рекомендации статьи с учетом параметра новизны с последующим выводом функции потерь для учета этого фактора.

Заключение

Роль персонализированных рекомендаций в процессе цифровизации всех областей интернета также становится все более важной, особенно в области новостей. Одной из важнейших задач сегодня является интеграции существующих персонализированных рекомендательных технологий в ин-

тернет-новости; другая не менее важная задача – построить оптимальную модель предпочтений пользователей, всесторонне учитывающую множество факторов. Решая эту важную задачу, появляется возможность создания новостной рекомендательной системы с исчерпывающей контекстуальной информацией. Изучение масштабируемых контекстно-зависимых методов рекомендаций имеет большое значение для продвижения в практических приложениях. До сих пор существуют две проблемы в исследовании контекстно-зависимых рекомендательных систем. Одна из проблем заключается в том, что каждая система рекомендаций в основном предназначена для конкретного сценария применения, что ограничивает масштабируемость данной системы. Другая проблема заключается в том, что все контекстно-зависимые рекомендательные системы содержат одинаковую контекстуальную информацию, но, несмотря на это, одна и та же контекстуальная информация может обладать различной структурой хранения и обработки, что ограничивает обмен данными между различными контекстно-зависимыми рекомендательными службами.

Список литературы

1. Bao J., Mokbel M.F., Chow C.Y. GeoFeed: A location-aware news feed system. Proc. IEEE Int. Conf. Data Eng. 2012. P. 54–65.
2. Chen C., Meng X., Xu Z., Lukasiewicz T. Location-aware personalized news recommendation with deep semantic analysis. IEEE Access. 2017. Vol. 5. P. 1624–1638.
3. Epure E.V., Kille B., Ingvaldsen J.E., Deneckere R., Salinesi C., Albayrak S. Recommending personalized news in short user sessions. In Proc. 11th ACM Conf. Recommender Syst. (RecSys). 2017. P. 121–129.
4. Yeung K.F., Yang Y., Ndzi D. A proactive personalised mobile recommendation system using analytic hierarchy process and Bayesian network. Journal of Internet Services and Applications. 2012. P. 195–214.
5. Клюквин Р.В., Белов Ю.С. Использование косинуса схожести в рекомендательных системах на основе коллаборативной фильтрации // Электронный журнал: наука, техника и образование. 2017. № 2 (12). С. 131–136.
6. Mohallick I., Özgöbek Ö. Exploring privacy concerns in news recommender systems. In Proc. Int. Conf. Web Intell. (WI). 2017. P. 1054–1061.
7. Gulla J.A., Marco C., Fidjestøl A.D., Ingvaldsen J.E., Özgöbek Ö. The intricacies of time in news recommendation. In Proc. 24th Conf. User Modeling, Adaptation Pers. (UMAP). 2016. P. 1–6.
8. Lommatzsch A., Kille B., Albayrak S. Incorporating context and trends in news recommender systems. In Proc. Int. Conf. Web Intell. (WI), 2017. P. 1062–1068.
9. Jugovac M., Jannach D., Lerche L. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. Expert Syst. Appl. 2017. Vol. 81. P. 321–331.