УДК 004.942

АРХИТЕКТУРА СИСТЕМЫ РЕКОМЕНДАЦИИ НОВОСТЕЙ, ОСНОВАННАЯ НА ПРИНЦИПЕ ПРОФИЛИРОВАНИЯ

Колебцев В.И., Белов Ю.С., Козина А.В.

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

Рекомендательные системы помогают пользователям справляться с информационной перегрузкой, предоставляя им индивидуальные предложения в различных областях (музыка, кино, товары в интернет-магазинах, новости и др.). Исходя из новостного контента и информации о пользователе, помощь пользователям новостных ресурсов в поиске интересных статей, максимально соответствующих их предпочтениям, стала одной из главных задач для современных интернет-порталов и мобильных приложений. В данной статье предлагается модель профиля пользователя для описания предпочтений пользователей с различных точек зрения. Персонализированные новостные рекомендации ориентированы на изучение связей между недавно опубликованными новостными статьями и профилями пользователей. Рекомендация новостей часто считается сложной задачей, поскольку релевантность статьи для пользователя может зависеть от множества факторов, включая краткосрочные интересы пользователя к чтению, контекст читателя, «свежесть» статьи или популярность статьи. Но еще большей сложностью представляется создание профиля пользователя, который должен постоянно адаптироваться под интересы человека, в зависимости от новостной повестки, настроения пользователя, местоположения и даже времени суток. Данная статья рассматривает одну из возможных структур пользовательского профиля, который может модифицироваться в зависимости от конкретной задачи.

Ключевые слова: рекомендация новостей, персонализация, метод профилирования пользователей

ARCHITECTURE OF THE NEWS RECOMMENDATION SYSTEM BASED ON THE PROFILING PRINCIPLE

Kolebtsev V.I., Belov Yu.S., Kozina A.V.

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

Recommendation systems help users cope with information overload by providing them with individual offers in various areas (music, movies, products in online stores, news, etc.). Based on news content and user information, helping users of news resources find interesting articles that best match their preferences has become one of the main tasks for modern Internet portals and mobile applications. This article offers a user profile model for describing user preferences from various points of view. Personalized news recommendations focus on exploring the links between recently published news articles and user profiles. News recommendation is often considered a difficult task, since the relevance of an article to the user may depend on many factors, including the user's short-term reading interests, the reader's context, the «freshness» of the article, or the popularity of the article. But it is even more difficult to create a user profile that must constantly adapt to the interests of the person, depending on the news agenda, the user's mood, location, and even the time of day. This article discusses one of the possible structures of a user profile, which can be modified depending on the specific task.

Keywords: news recommendation, personalization, user profiling method

Рекомендательные системы в настоящее время широко используются в различных онлайн-сервисах, в которых они помогают пользователям находить релевантный контент [1]. Сегодня области применения рекомендательных систем варьируются от предложения товаров на сайтах магазинов и музыкальных рекомендаций на потоковых платформах до рекомендаций друзей в социальных сетях [2].

Одной из самых ранних областей применения является рекомендация онлайн-новостей. Новостная рекомендация иногда рассматривается как особенно сложная, так как она имеет ряд отличительных особенностей.

Первая проблема заключается в том, что система часто не может полагаться на долгосрочные профили предпочтений пользователей [3]. Как правило, большинство пользователей не входят в систему, и их краткосрочные интересы в новостных статьях должны оцениваться только по нескольким зарегистрированным взаимодействиям (как правило? это клики по заголовкам или проведенное за чтением статьи время), что приводит к проблеме рекомендаций на основе сеанса. В последние годы наблюдается повышенный интерес к проблеме сессионных рекомендаций [4], когда задача состоит в том, чтобы рекомендовать соответствующие элементы с учетом текущей сессии пользователя.

Вторая проблема заключается в том, что во многих системах рекомендаций новостей профили пользователей являются односторонними, а это в свою очередь приводит к построению однобокой модели, что никак не отражает реальные предпочтения пользователя.

Еще один из факторов, который негативно влияет на рекомендацию новостей, то, что предпочтения пользователей в отношении новостей совершенно разные. При построении краткосрочного профиля большинство исследований не рассматривают относительно ранние записи просмотров или используют только несколько последних записей просмотра. Это может привести к многочисленным непредвиденным обстоятельствам и неправильному пониманию предпочтений пользователя, или результаты рекомендаций будут слишком похожи на то, что пользователь только что прочитал.

Для решения упомянутых выше проблем целесообразно использовать систему рекомендаций новостей, которая расширяет профиль пользователя до трех компонент: сбор и обработка новостей, метод профилирования пользователей и персонализированная рекомендация новостей.

Цель исследования: рассмотреть этапы профилирования пользователя и новостной статьи в новостной системе рекомендаций. Выявить основные этапы и структуру каждого этапа профилирования. Определить основные сущности, которые входят в модель профиля, и представить их в виде векторного пространства.

Общие принципы построения рекомендательной системы

Перед тем как перейти к подробному рассмотрению структуры системы рекомендаций на основе профилирования, рассмотрим некоторые общие принципы, с их достоинствами и недостатками, для построения рекомендательной системы.

Обычно новостной контент представляется моделью векторного пространства (например, TF-IDF) или тематическими распределениями, полученными с помощью языковой модели (например, PLSI и LDA), и оценивает отношения между новостными статьями с помощью конкретного метода измерения сходства. Например, News Dude был персонализированной рекомендацией, которая сочетает TF-IDF с алгоритмом K-ближайшего соседа для рекомендации новостей. MONERS [5] – это мобильная система рекомендаций новостей в интернете, которая включает в себя атрибуты новостных статей и пользовательские предпочтения в отношении категорий и новостных статей в процессе моделирования. Такие системы рекомендаций обычно легко реализуемы. Однако в некоторых сценариях профиль пользователя с bag-of-words недостаточно точно отражает предпочтения пользователя.

На практике большинство систем основываются на поведении пользователя в истории рейтингов и используют набор похожих пользователей для прогнозирования предпочтения в новостях или моделируют поведение пользователя вероятностным образом. В случае высокого сходства исторического поведения пользователей система могла бы эффективно улавливать предпочтения пользователей. Однако многие пользователи не имеют достаточную историю их поведения на новостном ресурсе или количество пользователей в системе недостаточно велико, этот недостаток известен как проблема холодного запуска.

Структура системы рекомендаций на основе профилирования

Рассмотрим краткую структуру системы рекомендаций, которая может дополняться и модифицироваться в зависимости от конкретной задачи и целей. Структура рекомендаций новостей, используемая в этой статье, разделена на следующие части: сбор и обработка новостей, профилирование пользователя, основанное на поведении чтения и популярности отдельных новостей, а также динамический персонализированный метод рекомендаций. Рассмотрим каждый этап подробно.

Сбор и обработка новостей (рис. 1): в данной статье представлена трехступенчатая профильная модель новостей. На первом этапе сканируется массив новостей (например, новости могут браться с основных новостных сайтов или тех источников, которые пользователь выбрал в случае регистрации) и извлекаются ключевые слова из новостей, чтобы создать векторную модель пространства. Затем анализируется распределение тем по языковой модели (например, LDA) для второго этапа. В итоге после завершения данного этапа формируется «профиль» новости.

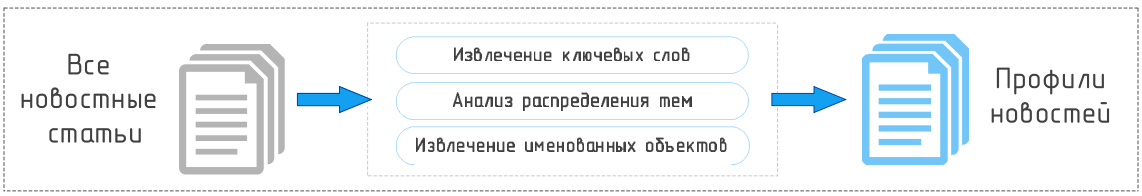

Рис. 1. Структура модуля «Сбор и обработка новостей»

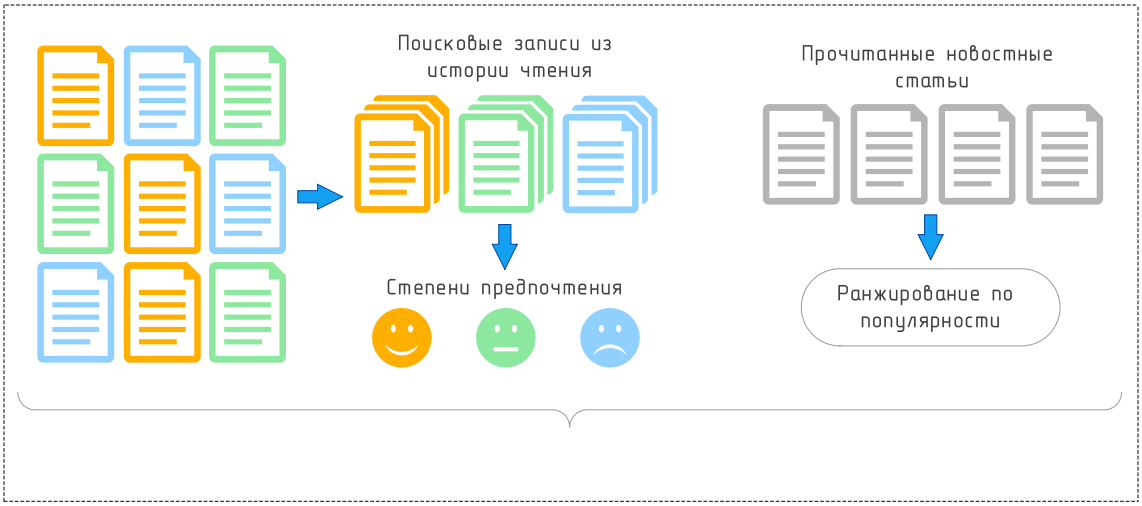

Рис. 2. Структура модуля «Профилирование пользователей»

Многоступенчатая модель [6] новостного профиля строится следующим образом. Во-первых, каждая новость должна пройти через систему сегментации для получения результатов сегментации слов. Затем строится векторная пространственная модель с ключевыми словами новостей по алгоритму TF-IDF, используя результаты сегментации. Извлекаются именованные сущности новостей. Наконец, мы получаем тематическое распределение новостей по LDA (латентное распределение Дирихле). Модель профиля каждой новости может быть выражена как

$$V_n = \langle F_n, G_n, E_n \rangle,$$

где F_n – вектор ключевых слов (чтобы получить ключевые слова, которые могут представлять новостную статью, вычисляется вес каждого слова в результатах сегментации по TF-IDF), G_n – вектор распределения тем (Читатели новостей также будут заинтересованы в новостях с аналогичными тематическими распределениями. Поэтому необходимо проанализировать тематику распространения новостей. В общем случае выявление скрытой темы текстового корпуса обычно проводится с использованием вероятностных языковых моделей, таких как PLSI и LDA, и путем извлечения списка репрезентативных слов из исходного корпуса вместе с соответствующими весами.), E_n – вектор именованных сущностей (Многие люди предпочитают читать новостные статьи с определенными именованными сущностями, в которых они заинтересованы. В качестве именованных сущностей могут выступать места, в которых происходит новостной сюжет, личность, которая участвует в сюжете. Поэтому необходимо учитывать и включать именованные сущности как часть новостного профиля.).

Профилирование пользователя (рис. 2): так же как и профиль новости, профиль пользователя состоит из трех этапов, которые основываются на истории чтения пользователя. На первом этапе представлены некоторые ключевые слова новостных статей, которые интересуют пользователя. Второй этап представляет собой тематическое распределение предпочтений пользователя [7]. Третий этап представляет собой именование сущностей, в которых заинтересован пользователь [8]. В данной статье поведение пользователя при чтении подразделяется на несколько типов (каждый из которых представляет собой различные степени предпочтения соответственно). Конкретная реализация классификации новостей может отличаться в зависимости от задачи.

Очевидно, что предпочтения пользователя меняются с течением времени. Поэтому в рекомендациях необходимо учитывать долгосрочные и краткосрочные предпочтения пользователей. Долгосрочные предпочтения пользователя относятся к общим предпочтениям, которые пользователь сформировал от использования системы до текущего времени. Краткосрочные предпочтения преимущественно основаны на недавнем поведении пользователя при чтении [9]. Основываясь на приведенном выше описании этапов профилирования, мы предлагаем построить профиль пользователя по трем различным, но взаимосвязанным элементам: ключевые слова новостей, именованные сущности новостей и распределение тем новостных статей.

Каждый профиль пользователя может быть выражен как

$$V_u = \langle F_u, G_u, E_u \rangle,$$

что соответствует трехступенчатой модели новостей. Однако есть некоторые отличия между этим выражением и профилем новостей, где F_u – вектор ключевых слов, собранных из новостных статей, к которым пользователь обращался в прошлом, где каждая запись состоит из репрезентативного слова, соответствующего веса и последнего времени, когда пользователь обращался к нему; G_u – вектор распределения тем, собранных из новостных статей, к которым пользователь обращался в прошлом, где каждая запись состоит из репрезентативного идентификатора темы, соответствующего веса и последнего времени обращения пользователя к ней; E_u – вектор именованных сущностей, собранных из новостных статей, к которым пользователь обращался в прошлом, где каждая запись состоит из репрезентативной именованной сущности, соответствующего веса и последнего времени обращения к ней пользователя.

Причина, по которой в системе используем как G_u, так и E_u, заключается в том, что G_u может отражать только интерес человека к новостным темам. Но предпочтения людей в отношении тем могут основываться на конкретном содержании. Например, некоторые люди предпочитают новости о музыке и кино, но они читают только новости о своей любимой звезде. Некоторым людям нравятся новости футбола, но они интересуются только командой «Спартак», а не «Локомотивом». И это именно то, что возможно учитывать, введя параметр F_u.

Рассмотрим один из возможных факторов, влияющий на степень предпочтения новости. Когда пользователи читают новости в интернете, степень предпочтения новостей и скорость чтения сильно коррелируются. Когда пользователь очень заинтересован в текущих новостях, он будет внимательно читать новости, и вследствие этого будет прочитана большая часть новостей. Таким образом, скорость чтения будет относительно медленной.

Однако, когда пользователь не заинтересован в текущих новостях, значительное количество контента будет пропущено. Пользователь может прочитать только несколько предложений, поэтому скорость чтения будет относительно высокой. Таким образом, скорость чтения может в значительной степени отражать степень предпочтения пользователя.

Когда пользователь читает новостные статьи разной длины, скорость чтения у него также будет разной. Специфика новостных статей заключается в том, что скорость чтения увеличивается с увеличением количества символов в статье. Когда пользователи читают длинные новости, даже если они заинтересованы в этом, скорость чтения относительно высока. Это происходит из-за того, что при чтении длинных новостей, много контента будет пропущено, так как большинство людей просто читают первую половину новостной статьи, чтобы знать, о чем говорят в новостях. После чего пользователи пропускают вторую половину новостной статьи. Именно по этой причине скорость чтения длинных новостных статей очень высока. Однако для коротких новостей, если пользователь очень заинтересован в тексте, он будет читать почти каждое слово, поэтому скорость чтения будет медленнее.

Динамическая персонализированная рекомендация новостей (рис. 3): как уже не раз упоминалось в этой статье, рекомендация новостей должна производиться, используя долгосрочные и краткосрочные предпочтения пользователей. Именно поэтому необходимо учитывать специальную функцию, чувствительную ко времени, для построения краткосрочного профиля путем корректировки долгосрочного профиля пользователя. После этого вычисляется сходство между каждой новостью и профилем пользователя. Затем происходит ранжирование новостей по сходству. Используя ранговые результаты долгосрочных и краткосрочных предпочтений, будут получены рекомендательные результирующие наборы долгосрочных и краткосрочных профилей. Окончательный результат рекомендации выбирается из долгосрочных и краткосрочных результатов рекомендации на основе соотношения, которое рассчитывается на основе истории просмотров пользователя. Наконец, нужно скорректировать соотношение выбора двух рекомендуемых наборов.

Предпочтения пользователей для долгосрочных и краткосрочных профилей могут варьироваться в зависимости от времени, и существуют различия между различными пользователями. Например, предпочтения некоторых пользователей стабильны, и их интересы и тематики в новостных статьях будут редко изменяться. Таким образом, для такого типа пользователей долгосрочный профиль подходит лучше. С другой стороны, предпочтения другой части пользователей изменчивы. Таким образом, рекомендательный результат краткосрочного профиля лучше соответствует их требованиям. Поэтому необходимо ввести персонализированный коэффициент отбора, основанный на исторических рекомендациях пользователя.

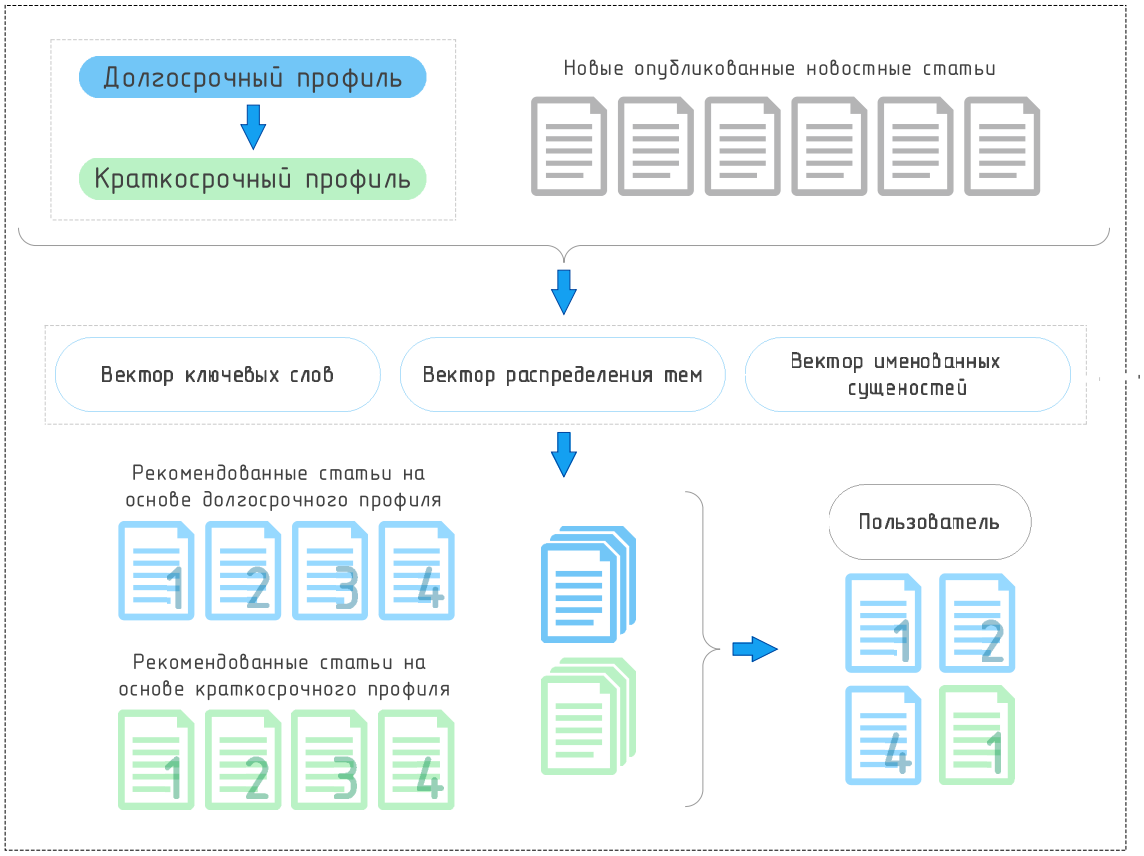

Рис. 3. Структура модуля «Динамическая персонализированная рекомендация»

Заключение

В этой статье представлен более персонализированный метод рекомендации новостей, основанный на профилировании. Профили пользователей строятся с учетом трех различных характеристик: ключевые слова новостей, распределение тем новостей и именованные сущности. Таким образом, пользователь, зарегистрированный в системе длительное время, будет получать качественные рекомендации новостных статей, так как его собственные краткосрочные и долгосрочные профили будут постоянно уточняться и модифицироваться друг относительно друга.

Список литературы

1. Lin C., Xie R., Guan X., Li L., Li T. Personalized news recommendation via implicit social experts. Inf. Sci. 2014. vol. 254. P. 1–18.

2. Lv P., Meng X., Zhang Y. FeRe: Exploiting influence of multidimensional features resided in news domain for recommendation. Inf. Process. Manage. 2017. vol. 53. no. 5. P. 1215–1241.

3. Saranya K.G., Sadasivam G.S. Personalized news article recommendation with novelty using collaborative filtering based rough set theory. Mobile Netw. Appl. 2017. vol. 22. no. 4. P. 719–729.

4. Ludewig M., Jannach D. Evaluation of session-based recommendation algorithms. User Model. User-Adapt. Interact. 2018. vol. 28. no. 4–5. P. 331–390.

5. Lee H.J., Park S.J. MONERS: A news recommender for the mobile Web. Expert Syst. Appl. 2007. vol. 32. no. 1. P. 143–150.

6. Fortuna B., Moore P., Grobelnik M. Interpreting news recommendation models. Proc. 24th Int. Conf. World Wide Web (WWW Companion). 2015. P. 891–892.

7. Quadrana M., Cremonesi P., Jannach D. Sequence-aware recommender systems. ACM Comput. Surv. 2018. vol. 51. no. 4. Art. no. 66.

8. Jannach D., Lerche L., Kamehkhosh I., Jugovac M. What recommenders recommend: An analysis of recommendation biases and possible countermeasures. User Model. User-Adapted Interact. 2015. vol. 25. no. 5. P. 427–491.

9. 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.