Сергеев В.В.
студент Московского государственного университета печати им. Ивана Федорова
Sergeyev V.V.

АЛГОРИТМ СБОРА ИНФОРМАЦИИ ПО ПРЕДПОЧТЕНИЯМ В СИСТЕМЕ АВТОМАТИЗИРОВАННОГО СБОРА ДАННЫХ

Аннотация. *В статье раскрывается алгоритм поиска и сбора информации в сети интернет в целях экономии времени и увеличения производительности труда человека.*

ALGORITHM OF GATHERING OF THE INFORMATION ON PREFERENCES IN SYSTEM OF THE AUTOMATED DATA GATHERING

SUMMARY. *In article the algorithm of search and information gathering in a network the Internet with a view of economy of time and increase in labor productivity of the person reveals.*

Ключевые слова: информационные технологии, RSS, пользовательская среда, база данных, поиск по ключевым словам, рейтинг, статья.

Keywords: Information technology, RSS, the user environment, a database, search in keywords, a rating, article.

Интернет становится все более и более ориентированным на конечного пользователя. В борьбе за пользователя выживают только самые необходимые, уникальные, профессиональные и, самое главное, удобные для использования системы. Главным ресурсом современного мира становится время. По этой причине появляется все больше и больше программных решений направленных на повышение производительности труда человека за счет экономии его времени. Человеку уже не обязательно самому ходить за покупками, посещать учебные заведения, путешествовать по всему миру, чтобы ознакомится с интересующими его достопримечательностями, нет необходимости покупать журналы, газеты, книги для того, чтобы быть в курсе всего происходящего в мире. Все это можно легко сделать, сидя у себя или же по дороге на работу, благодаря бурному развитию информационных технологий.

Для того, чтобы быть в курсе всех мировых новостей, новостей своего региона или же населенного пункта, как правило, недостаточно использовать один информационный портал, на котором можно будет находить всю интересующую вас информацию. Не всегда качество предоставляемой конкретным информационным ресурсом информации соответствует требованию и пожеланию конкретного пользователя.

Казалось, что решение данной проблемы должно было прийти с появлением RSS — специального XML-формата, разработанного для описания лент

новостей, анонсов статей, изменений в блогах и т. п. Несмотря на некоторые различия между разными версиями, RSS стал очень удобным инструментом для передачи краткой информации об изменении информации интернет-ресурса. Для удобства чтения RSS подписки конечного пользователя появились специальные программы-агрегаторы. Самые популярные — Google Reader, Yandex.lenta и Netvibes.

Можно довольствоваться RSS-подпиской, но RSS не может предугадать, понравится ли конкретная статья читателю, и не в состоянии оценить, насколько она его заинтересует и заинтересует ли вообще. RSS показывает наличие новостей, но не может порекомендовать ту или иную статью.

Разрабатываемый алгоритм, внедренный в автоматизированную систему сбора новостных статей, направлен на решение данной проблемы. Описание системы, реализующей данный алгоритм, и сам алгоритм приведены ниже (рис. 1).

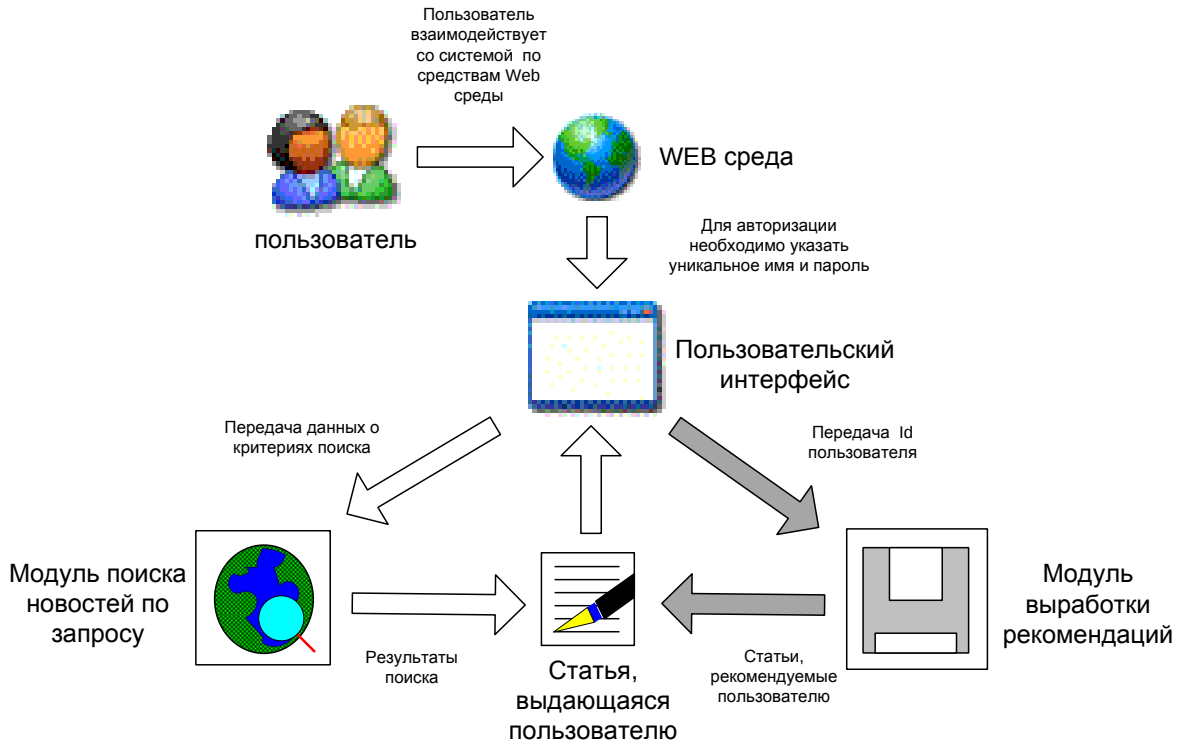

Рисунок 1. Общая структурная схема системы реализующей алгоритм выработки рекомендаций

Главная задача заключается в объединении в одной системе всех возможных новостных ресурсов в одном месте.

Реализация автоматизированной системы построена на использовании новостных RSS-лент, имеющихся практически на любом сайте и выступающих в данном случае как основной ресурс информации.

При входе на сайт пользователю будет предложено зарегистрироваться в системе, указав уникальное имя и пароль, или, если пользователь уже зарегистрирован, будет предложено войти под зарегистрированной учетной записью. При входе под своей учетной записью человек попадет в пользовательскую среду, в которой будет предложено воспользоваться настройками фильтров поиска интересующих его новостей по разнообразным критериям. К таким

критериям относятся: источники полученной информации, время публикации, рубрика, а также наличие или отсутствие ключевых слов в публикациях.

Изначально в системе будет доступен список источников RSS-рассылок [S], которые могут быть выбраны для поиска интересующей информации. При этом будет реализована возможность добавления своих источников с RSS-данными. После выбора источников данных, нам будет предложен выбор диапазона времени, в течение которого мы будем получать новости [T=T1-T2]. Сортировка по времени позволяет просматривать новости, сохранившиеся в базе данных, за произвольный момент времени, ориентируясь по датам появления публикаций. Как известно, каждый человек имеет свое мнение и свою манеру описания событий, поэтому возможность сортировки по автору [A] статьи даст возможность получателю информации более тщательно подходить к выбору источников данных.

Результат поиска W(S,T,A) = W(S)xW(T)xW(A)

На главной странице читателю будет предоставлен список автоматически подобранных статей с учетом настроек фильтров. При желании можно будет воспользоваться поиском статей по ключевым словам [L]. При поиске по ключевым словам учитываются частотные характеристики появления в тексте заданных слов [N].

Простой подсчет количества ключевых слов в статье не может являться верной оценкой верификации. Подводные камни при использовании подсчета количества ключевых слов в статье нас ожидают, когда сравниваются различные по объему тексты. Само собой разумеется, что в более объемном тексте вероятность присутствия заданного слова выше пропорционально объему статьи. Поэтому используются условные частоты присутствия ключевых слов. Условные частоты считаются по формуле pr(c,t)=f(c,t)/L(t), где L(t) – длина текста t, а f(c,t) – частота употребления в статье заданных слов.

Результат поиска W =W(S,T,A)xW(pr) становится более точным.

Очень важным моментом взаимодействия с пользователем является обратная связь с конечным читателем (рис. 2).

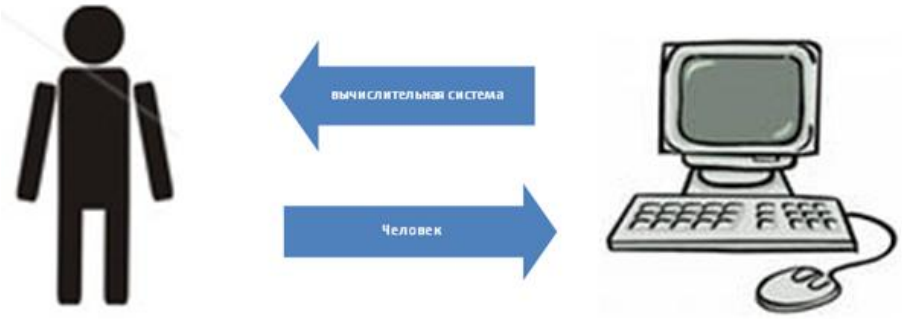

Рисунок 2. Обратная связь с пользователем

Именно обратная связь позволит выявлять интересы пользователей, их приоритеты и позволяет адаптировать контент под интересы конечного пользователя.

Для этих целей используется система рейтинга по каждой статье [Ri]. После прочтения статьи пользователю будет предложено оценить актуальность статьи по определенной заданной шкале оценок. Таким образом, мы будем получать рейтинг каждой прочтенной статьи по мнению конкретного читателя. Постепенно в системе будут накапливаться данные о предпочтениях каждого зарегистрированного пользователя. Эти данные позволяют определить своеобразное «облако интересов» конкретного пользователя.

Фактически можно подсчитывать средний рейтинг по каждой статье, основываясь на средних оценках всех пользователей

оценка = (оценка1+оценка2+оценка3+…+оценкаN)/N, где N – количество проставленных оценок.

Однако, надо заметить, что все читатели отличаются друг от друга. А статьи могут быть направлены на разные социальные группы и поэтому могут получить совершенно противоположные оценки в разных пользовательских кластерах. Поэтому критерии подбора статей по максимальной средней оценке не является оптимальным.

Пользователю необходимо предоставить возможность получения мнений о статье людей социальной группы, максимально соответствующей критериям поиска. [illegible]. Для этого, основываясь на данных о прочтенных пользователем статьях и на основе проставленных статье оценок, необходимо выделить группы пользователей с максимально общими интересами.

Система, основываясь на оценки, проставленные пользователем статьям, будет собирать некий психологический портрет пользователя. Собранную информацию можно использовать для кластеризации пользователей по предпочтениям. Для этих целей будет использоваться коэффициент корреляции Пирсона, который находится по формуле [illegible]

где M – математическое ожидание дискретного распределения

$$M[X]=\sum_{i=1}^{\infty} x_i p_i$$

Если 2 человека попали в одну группу, то можно предположить, что статьи, прочитанные одним человеком, с определенной вероятностью понравятся другому человеку, попавшую в одну с ним группу. Соответственно, система может рекомендовать пользователю прочитать ту или иную статью (рис. 3).

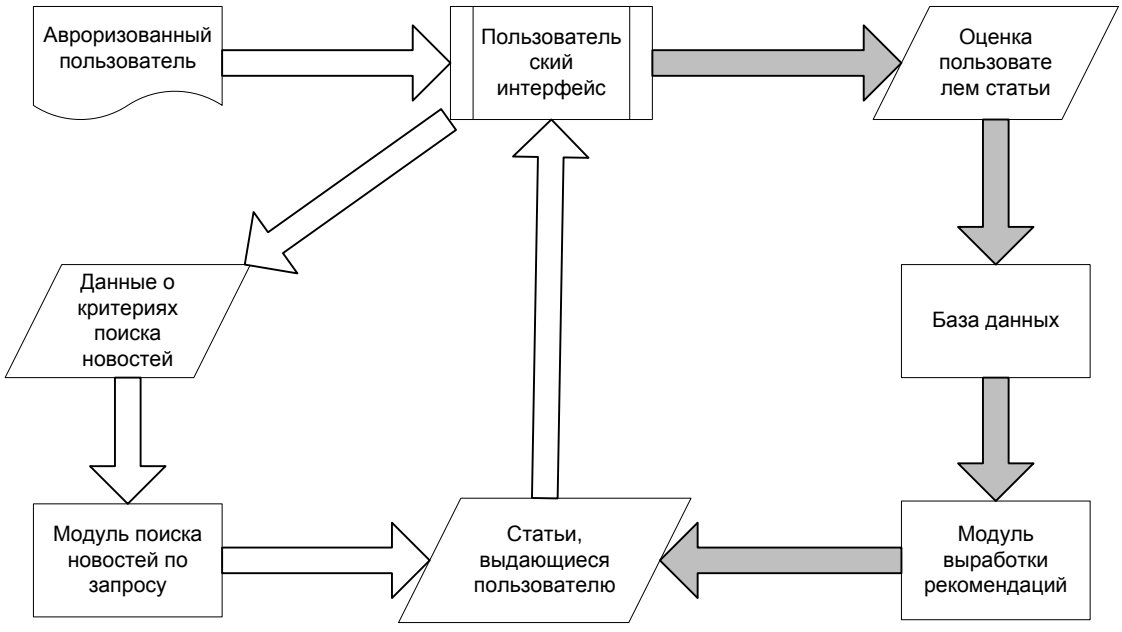

Рисунок 3. Серым цветом условно показан процесс выработки рекомендаций пользователю на основе проставленных им оценок.

Имея данные об оценке конкретной статьи различными кластерами пользователей, мы можем получить дифференцированную картину интересов каждой социальной группы с определенной дисперсией, обусловленной индивидуальными особенностями пользователей $D(X) = M[X - M(X)]^2$, где X – некоторая произвольная оценка статьи и X принадлежит пространству всех оценок $X \in \mathrm{v}$. Однако даже эта дисперсия на выходе даст положительный эффект, так как поможет составить более полную картину востребованности информации среди социальных групп, а также в будущем эти данные могут быть использованы для выделения нового кластера.

Во всех областях есть эксперты, достаточно хорошо разбирающиеся в той или иной области. Но не каждый пользователь хочет читать только то, что читали люди с интересами, схожими с его интересами. Пользователю будет дана возможность выбрать себе несколько экспертов, на оценки которых он сможет ориентироваться при выборе той или иной статьи, мнение которых о статьях в определенной рубрике ему наиболее важно.

Таким образом, у нас получается множество пользователей с неким подмножеством интересов, а каждый из интересов, в свою очередь, имеет подмножество характеристик, таких как приоритет, уровень (эксперт, любитель, новичок).

После каждой статьи необходимо реализовать возможность добавления комментария, отзыва или рецензии всем зарегистрированным пользователям. Если читателю понравится рецензия, то он сможет добавить к рейтингу автора комментария несколько условных баллов, а если не понравится, то вычесть несколько условных баллов. Таким образом, у рецензента складывается рейтинг. Рейтинг будет 2-х видов: обобщенный и персональный. Обобщенный рейтинг отображает среднюю оценку читателей данного рецензента, а персональный будет отображать персональную оценку читателя. Расширяя понятие «рейтинг рецензента», можно будет вывести рейтинг рецензента касательно конкретной рубрики. Система накапливает данные о ваших оценках рецензентам и на основе статистических данных обучается понимать, кому вы доверяете свое мнение в той или иной рубрике. Фактически данный человек является для вас экспертом в данной области. Тогда с большой уверенностью можно сказать, что, если статья понравилась человеку, мнение которого вы разделяете касательно данной тематики статьи, то с высокой вероятностью она понравится и вам. Таким образом, мы получаем возможность интеллектуального подбора статей для конечного пользователя. Точность верификации будет достигаться постепенно за счет увеличения статистических данных о пользователе. На первом этапе пользователю будут доступна возможность выбора статей на основе статистики просмотра и оценки других пользователей.

Объединение алгоритма подбора статей, основанного на кластеризации читателей по проставленным ими оценкам, и алгоритма подбора статей, основанного на выборе статей высоко оцененных критиками, обладающими максимальным рейтингом доверия, приводит к повышению вероятности заинтересованности пользователя в рекомендуемых системой статьях (рис. 4).

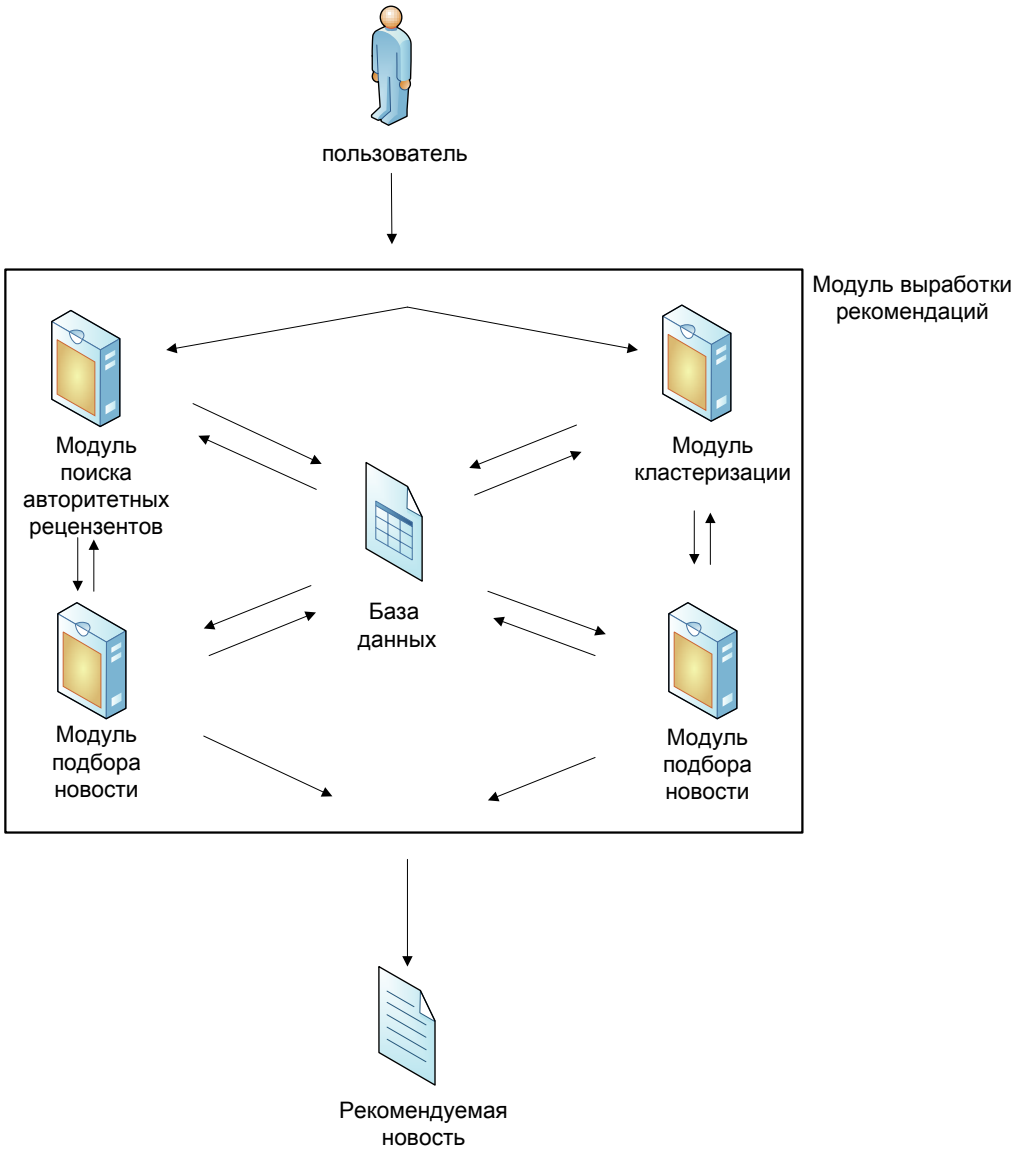

Рисунок 4. Структурная схема работы модуля выработки рекомендаций

В результате мы получаем алгоритм, позволяющий активному пользователю новостного ресурса автоматически получать статьи, максимально совпадающие с его интересом. Это позволяет сэкономить время на просмотр всего потока новостной информации, что в целом способствует увеличению производительности труда человека.

ЛИТЕРАТУРА

1. Михайлов, А.Г.Проектирование информационных систем в Internet. Руководство для менеджера[Текст] / А.Г. Михайлов.– М. : Информ-Знание, 2000. – 116 с.
2. Бушуева, Л.И. Роль Интернет-услуг в практической маркетинговой деятельности [Текст] / Л.И.Бушуева // Маркетинг в России и за рубежом. – 2001. – № 4. – С. 67-82.
3. Алексеев, А.А. Маркетинговые исследования рынка услуг [Электронный ресурс] / А.А. Алексеев. – Электрон. дан. – М., 1998 – . – Режим доступа : http: // www.marketing.spb.ru/read/m17/index.htm. – Загл. с экрана.