

Рекомендательный сервис на основе Википедии и социальной сети

*И.В.Браиловский, Л.В.Дмитриев, Н.Е.Иванов,
Н.Ю.Комаров, Д.С.Уманский, И.В.Фомин, А.В.Чечендаев,
ф-т ВМК МГУ им. М.В. Ломоносова*

В работе описан рекомендательный сервис, разрабатываемый коллективом авторов в лаборатории Открытых информационных технологий факультета ВМК МГУ, а также проведен обзор технологий и средств социального поиска.

1. Введение

Проблема классификации информации и информационного поиска в целом актуальна как никогда: количество производимой информации на цифровых носителях увеличивается экспоненциально, в то время как способности человека к восприятию и переработке информации растут незначительно, или даже остаются на том же уровне.

В настоящее время наполнение глобальной сети Интернет значительно выросло в связи с повышением ее доступности для широких масс, а так же появлению проектов Веб 2.0, наиболее яркими примерами которых являются блоги, социальные сети и Википедия. В таких проектах основным наполнением является информация, созданная и внесенная на сайт непосредственно пользователями данного ресурса. В результате суммарное количество информации в Сети растет, а ценность ее в среднем уменьшается, что приводит к тому, что пользователю становится все труднее как ориентироваться в уже известных ему ресурсах, так и находить новые источники интересной информации.

Одна из главных проблем в Интернете – получение доступа к информации, заслуживающей доверия. Например, в общем трафике электронной почты доля спама и вирусов достигает уровня около 90 процентов. Проблема спама, скрытой рекламы и бессмысленной информации стоит остро и на информационных сайтах. При современном уровне развития технологий с помощью автоматических методов можно отсеять лишь некоторые классы бесполезной и вредной информации.

В данной работе предлагается способ построения рекомендательного сервиса, основанного на персонифицированном рейтинговании информации – то есть рейтинговании с учетом только голосов пользователей, имеющих социальную связь (связь в социальной сети) с запрашивающим информацию.

2. Социальный поиск и структурирование информации

Развитие Сети привело к возникновению новых технических и социальных эффектов, которые в ставшей классической статье [9] (имеется русский перевод [16]) обобщил Тим О'Рейли под общим названием Веб 2.0. Среди этих изменений и переход от иерархической (таксономии) к горизонтальной (фолксономии) структуре связей. В настоящее время теги применяются для категоризации информации не только на сайтах с пользовательским контентом, причем в небольших тематических ресурсах с контролируемыми источниками информации возможно ручное поддержание системы тегов в консистентном состоянии, обеспечивающем качественную навигацию по тематическим блокам [15]. Социализация Веба продолжается, и теги все шире используются для навигации [3]. Пользуются популярностью и сервисы социальных закладок, являющиеся развитием идеи локального хранения закладок. Среди таких сервисов следует отметить Delicious [4] и БобрДобр [12]. На социальных новостных сайтах для оценки новостей применяется рейтингование, как, например, на Digg [5] и ХабраХабре [18].

Социальный поиск (social search) является перспективным и развивающимся направлением информационного поиска, среди перспективных направлений развития которого можно выделить collective social search, friend-filtered social search и collaborative search [6] (мы приводим англоязычную терминологию из-за отсутствия устоявшегося перевода на русский язык).

В поисковой системе с элементами социальной сети Wikia Search пользователи могли рейтинговать документы. Весной 2009 года проект был закрыт из-за малой посещаемости.

Весьма интересным является сервис Aardvark [1]. На основе указываемой пользователями при регистрации информации о своих областях знания и получаемой из социальной сети Facebook информации о связях между ними происходит переправление вопроса к знакомым, разбирающимся в соответствующей области.

Новым направлением в области социализации Веба являются рекомендательные сервисы (сервисы для предсказания насколько интересным пользователю будут некие информация или объекты на основе информации о нем). Наиболее близким к разрабатываемому является сервис Имхонет [14], идея которого была предложена Александром Долгиным [13]. Одной из целей системы Shopping2 [11] является построение графа потребления одежды.

Значительное количество информации не представлено в Сети в явном виде, а является экспертным знанием, в связи с чем активно развиваются технологии поиска экспертов [2, 8].

3. Описание сервиса

Создаваемый рекомендательный сервис Snewi (название образовано от Social Network Wiki) включает в себя черты социальной сети, сервиса закладок, коллективного блога и открытого каталога с некоторыми дополнительными механизмами оценки и сортировки информации. Основная идея заключается в том, что пользователь сам выбирает тех пользователей, чьи оценки должны учитываться при сортировке записей. Такой подход имеет ряд преимуществ перед традиционными.

- Сервис максимально защищен от спама. Спамеры могут создать большое количество фиктивных аккаунтов и бесконечно увеличивать обычный, персонифицированный рейтинг записи, но у обычного пользователя нет причин добавлять подобные аккаунты в свой список доверенных источников, поэтому при сортировке бессмысленные и рекламные записи окажутся в самом низу списка и не будут видны.
- Сервис максимально персонифицирован и не навязывает мнений. На традиционных сайтах обычно образуется некое большинство, обладающее более-менее известным мнением по многим вопросам. В данном сервисе порядок вывода записей зависит от мнения не большинства, а тех людей, которых пользователь выбирает сам, поэтому пользователь, мнение которого по определенному острому вопросу является нестандартным, сможет общаться с людьми, разделяющими его.

Интерфейс создаваемого сервиса имеет много общего с Википедией. Так же как в Википедии пользователь может ввести слово и получить статью. Так же как в Википедии он может использовать категории для навигации. Дополнительно есть несколько вкладок, в которых пользователь видит список ресурсов по данной статье, и они отсортированы по рейтингу.

Любой пользователь может добавить ресурс к статье. Ожидается, что ознакомившись с прикрепленными ресурсами пользователи будут давать им оценку. По сути, в Snewi используется система тегирования, в которой тегом является статья Википедии (набор тегов расширяем).

Для вычисления рейтинга используются оценки людей, которых выбрал сам пользователь (в терминах Snewi – советники). Для увеличения количества используемых оценок применяется рекурсия: используются советники советников и далее по графу социальных связей.

Отношение советника похоже на понятие друга или связи в современных социальных сетях, но более конкретно. Если пользователь назначает кого-то своим советником по теме, это значит, он считает его экспертом в определенной области и доверяет его оценкам по соответствующей теме. Каждый пользователь самостоятельно выбирает для

себя советников среди других пользователей системы. Таким образом, нет никакой иерархии советников – есть социальная сеть, в которой пользователи являются советниками друг для друга.

Выбирая одну из соответствующей статье вкладок, пользователь выбирает тип получаемой информации. Перечислим возможные типы.

- **Snawi Directory** – это список URL. Ссылка может указывать на сайт, конкретную статью, сообщение в блоге, файл с видео и любую другую доступную в Сети информацию.
- **Snawi Catalog** – это перечень коммерческих предложения обо всем, что продается за деньги (товары, услуги). Любой пользователь может внести свою услугу в этот раздел, а пользователю доступен список, отсортированный по оценкам советников.
- **Snawi News** – это новости, любая быстро устаревающая информация. Новостями могут быть как глобальные новости, выпущенные средством массовой информации, так и локальные новости, такие как предложение сходить в поход.

В информационных вкладках можно выбрать один из трех вариантов представления ресурсов.

- Только эта тема. Показываются ресурсы только по выбранной теме (статье). Очевидно, что по случайно выбранной статье Википедии шанс найти ресурсы, оцененные друзьями пользователя (советниками) невелик. Тогда используется технология облака советников, когда оценки рекурсивно ищутся у друзей друзей по графу социальных связей. У обычного пользователя социальной сети связи третьего порядка включают многие тысячи людей, и среди них может найтись тот, кто тоже интересуется этим вопросом.
- Данная категория. Если выбранная статья Википедии является категорией, то показываются ресурсы всех статей, которые попадают в эту категорию.
- Похожие ресурсы. Если данная статья не является категорией, то показываются ресурсы семантически близких статей.

Очевидно, что по некоторым категориям может быть найдено чудовищное количество ресурсов. С ростом количества ресурсов растет и количество оценок, что повышает качество сортировки.

Во вкладке **Advisers** можно настраивать советников по выбранной категории. Для оценки всех ресурсов, относящихся к статьям данной категории, будут использоваться голоса выбранных советников.

Можно назначить человека советником сразу по всем темам, но можно и уточнить его специализацию, то есть указать по каким темам он является советником. Тогда по другим темам его мнение учитываться не будет. Типичной является ситуация, когда пользователь, интересу-

ьясь какой-либо темой, например историей, выбирает себе советников по этой теме. При этом они не являются советниками по другим областям интереса, например музыке.

Советник по некоторой узкой теме, однако, обладает одним замечательным свойством. Он, как и любой человек, пользуется продуктами и услугами, и его мнение о них можно использовать. Поэтому при выборе продукта в сервисе Shewi можно назначить советниками уважаемых людей, которых знаешь, которым доверяешь и таким образом расширить облако советников. Так среди советников могут оказаться люди, у которых неудобно было бы спрашивать совета напрямую.

Облаком советников данного пользователя мы будем называть всех людей, мнение которых учитывается при составлении рейтинга для данного пользователя.

В облако советников входят, конечно, те, кого пользователь явно назначил советником, однако используется и рекурсивный обход графа социальных связей – возможен учет мнения советников советников, или даже советников советников советников. Глубина охвата задается пользователем или выбирается автоматически. Если пользователь выбрал тему, по которой мало ресурсов оцененных советниками первого уровня, то будет произведен поиск оценок советников советников и так далее до достижения необходимого количества ресурсов.

Облако советников строится для каждого конкретного пользователя для выбранной темы. Если пользователь ищет информацию по другой теме, то облако советников будет другим.

Если пользователь объявил себя открытым советником по некоторой теме, значит, его можно назначать своим советником без запроса авторизации. Этот человек просто делится своим мнением со всеми кого оно интересует. Мы полагаем, что открытость советника является типичной ситуацией, однако по соображениям приватности существует возможность для пользователя разрешать использование своих оценок только по специальному запросу.

В разрабатываемой системе информация запрашивается, а не проталкивается. Хорошо известен эффект вирусобразного распространения сообщений, когда сообщение пересылается по цепочке от одного пользователя к другому и так охватывает большое сообщество. Такой эффект наблюдается и в социальных сетях и при использовании электронной почты или систем мгновенного обмена сообщениями. С одной стороны, это замечательный механизм, который позволяет интересным и полезным сообщениям распространяться. При этом пользователь, направляющий письмо должен тем или иным способом предугадать положительную реакцию получателя на это письмо, а это сложно. Поэтому данный механизм во многих случаях не работает эффективно, то

есть приводит к потоку бесполезных (или даже вредных, как, например, письма счастья) для получателей сообщений.

В Snewi вирусобразное распространение тоже возможно, но по другому принципу. Пользователь голосует за ресурс (например, сообщение) и его видимость повышается для всех, у кого данный пользователь попадает в облако советников. Если у кого-то в облаке советников за сообщение проголосовало несколько человек, то его рейтинг увеличится и оно поднимется выше в списке ресурсов, что приведет к увеличению его эффективной области видимости. Информация только тогда попадает к пользователю, когда он ее явно запрашивает.

Выбор используемых ключевых слов является классической проблемой человеко-машинного взаимодействия [7], и в разрабатываемом сервисе предлагается решение на базе collaborative tagging с использованием информации из Википедии.

Все ресурсы, циркулирующие в Snewi должны иметь тему (tag). Он устанавливается тем пользователем, который добавляет этот ресурс в систему.

В Snewi в качестве тегов используются статьи Википедии. Это обеспечивает широкое покрытие тегами, близкое структуре интересов пользователей Сети. Другими преимуществами этого решения являются: возможность унификации написания важных тегов (отсутствие мусорных тегов с опечатками), учет синонимов (в Википедии такие статьи разделены), использование иерархии тем и категорий. При анализе статей Википедии (в том числе гиперссылок между ними) возможно определение семантической близости тем [17], что обеспечивает возможность поиска ресурсов по сходству.

Для классификации ресурсов используется технология collaborative tagging с контролируемым словарем. При этом в качестве такого словаря используются статьи Википедии. Возможно и добавление тега пользователем, как возможно и создание новой статьи автором Википедии.

Википедия содержит не только огромное количество статей из самых различных областей человеческой деятельности, но и информацию о связях между ними. В частности, механизм категорий позволяет определять, что один тег является частным случаем другого и использовать это для классификации и при поиске информации. А тексты статей позволяют оценивать семантическую близость понятий и, следовательно, тегов, что открывает еще одно направление использования и совершенствования системы.

4. Заключение

Предложенная архитектура имеет следующие особенности и преимущества. Отсутствует единый рейтинг ресурсов, одинаковый для всех

пользователей. Вместо этого рейтинг ресурсов считается индивидуальным для каждого пользователя, вес голоса зависит от близости по социальной сети проголосовавшего к просматриваемому списку ресурсов пользователю. Таким образом, частично решается проблема субъективности понятия качественной информации. Пользователь, назначая советников (устанавливая с ними социальную связь), доверяет им оценку информации и таким неявным образом определяет свои вкусы и предпочтения.

Использование социальной сети для оценки и фильтрации информации потенциально уменьшает проблему спама, значительно снижает возможности скрытой рекламы.

Создание альфа-версии позволило выявить слабые места проекта и улучшить концепцию. В настоящее время в лаборатории открытых информационных технологий факультета ВМиК МГУ ведется реализация прототипа сервиса, по мере которой возможно внесение изменений в его архитектуру. Перспективно обращение через соответствующие интерфейсы к пользовательской базе существующих социальных сетей, так как сложно мотивировать людей регистрироваться в еще одной. Еще одним перспективным направлением развития проекта является автоматизация поиска экспертов (открытых советников), имеющих общие интересы и взгляды.

Некоторые возможности прототипа рекомендательного сервиса Snewi доступны в сети Интернет по адресу snewi.org [10].

Литература

1. Aardwark. <http://vark.com/>
2. Balog K. People Search in the Enterprise. <http://staff.science.uva.nl/~kbalog/phd-thesis/>
3. Bielenberg K., Zacher M. Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation. <http://bielenberg.info/thesis.pdf>
4. Delicious. <http://delicious.com/>
5. Digg. <http://digg.com/>
6. Evans B. 3 Flavors of Social Search: What to Expect. http://www.readwriteweb.com/archives/3_flavors_of_social_search_what_to_expect.php
7. Furnas G. W., Landauer T. K., Gomez L. M., Dumais S. T. The Vocabulary Problem in Human-System Communication. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.103.8364>
8. Maybury M. T. Expert Finding Systems. http://mitre.org/work/tech_papers/tech_papers_06/06_1115/06_1115.pdf
9. O'Reilly T. What Is Web 2.0. <http://oreilly.com/web2/archive/what-is-web-20.html>
10. Snewi. <http://snewi.org/>
11. Shopping2. <http://shopping2.ru/>

12. БобрДобр. <http://bobrdobr.ru/>
13. Долгин А.Б. Экономика символического обмена. М.: Инфра-М, 2006. 632 с.
14. Имхонет. <http://imhonet.ru/>
15. Навоша Д. Теги и SEO для СМИ. <http://roem.ru/2008/11/28/navosha01/>
16. О'Рейли Т. Что такое Веб 2.0. <http://www.computerra.ru/think/234100/>
17. Турдаков Д. Измерение семантической близости концепций Википедии, основанное на анализе ссылок между статьями. <http://synthesis.ipi.ac.ru/sigmod/seminar/s20080327>
18. ХабраХабр. <http://habrahabr.ru/>