ISSN 2311-8814

«НАУКА | RASTUDENT.RU»

Электронный научно-практический журнал

График выхода: ежемесячно
Языки: русский, английский
ISSN: 2311-8814
Издатель: компания INFLASH
Учредитель: ИП Соколова А.С.
Место издания: г. Уфа, Российская Федерация
Прием статей по e-mail: rastudent@yandex.ru
Место издания: г. Уфа, Российская Федерация

Селивёрстов Е.В. Повышение качества рекомендательных систем за счет учета структуры документов // Наука-RASTUDENT.RU. – 2014. – No. 4(04-2014) / [Электронный ресурс] – Режим доступа. – URL: http://nauka-rastudent.ru/4/1312/

УДК 004

Селивёрстов Евгений Владимирович,
Магистрант
Кафедра Вычислительной Техники, Факультет ЭСТ
Московский Государственный Университет Леса
Мытищи, Московская область, Россия

Повышение качества рекомендательных систем за счет учета структуры документов

Аннотация: В данной статье пойдёт речь про работу рекомендательных систем, работающих с электронными документами. Описан принцип функционирования большинства рекомендательных систем, а также изложена мысль улучшения их результатов, за счет учета особенности структуры документов, с которыми будет работать система. В статье так же содержатся примеры результатов работы, продемонстрированные на реальном материале.
Ключевые слова: рекомендательные системы, рекомендательные алгоритмы, электронные документы, учет структуры документов.

Improving quality of recommender systems by accounting structure of documents

Seliverstov Evgeney Vladimirovich,
Master student
Department of Computer Engineering, Faculty of Electronics and Systems
Moscow State Forest University
Mytischi, Moscow region, Russia

Abstract: This article will talk about the work of recommendation systems, working with electronic documents. Describes the principle of functioning of the majority recommendation systems as well as set forth the idea of improving their results, taking into account the specifics of the structure of documents, which will operate the system. The article also contains examples of the results of recommendation algorithm, demonstrated in the real data.
Keywords: recommender systems, recommendation algorithms, electronic documents, accounting structure of documents.

С развитием информационных технологий и интернета человечество создало много новых областей деятельности, а также, оптимизировалось огромное количество старых. В качестве примера можно привести область,

связанную с созданием, обработкой, передачей и хранением документов. Если раньше всё это производилось вручную и на физических носителях (то есть на бумажных), то сейчас обслуживается с помощью компьютеров. Компьютеры способны хранить огромное количество документов, осуществлять по ним поиск за считанные секунды, копировать их, а также имеется возможность актуальную держать копию всей базы на другом компьютере, расположенном физически на другом континенте, на случай пожара или другого форс-мажора.

Электронные документы используются во многих системах. Это, например, сайты с рефератами и диссертациями, файловые обменники, интернет-библиотеки, журналы и т.д. В новостных и информационных порталах публикуемые статьи можно так же расценивать как электронный документ. Для повышения эффективности работы пользователей с этими системами, их авторы могут использовать различные средства, одним из которых может быть внедрение рекомендательного движка.

Рекомендательная система должна ненавязчиво предлагать пользователю документы, которые по её мнению должны оказаться интересными для него. Рекомендовать можно по-разному. Можно рекомендовать определенный набор документов для всех пользователей, можно рекомендовать самые часто скачиваемые документы, можно рекомендовать документы, похожие на те, что пользователь уже просматривал. Очевидно, что из всех этих методов последний является наиболее эффективным, а также персонализированным, то есть уникальным для каждого пользователя.

Большая часть рекомендательных систем использует модель векторного пространства [1]. В этой модели каждый документ представлен в виде вектора в n-мерном пространстве, где каждому измерению соответствует слово из этого документа. Значение измерения это вес слова, который указывает на степень связи слова с документом. Для определения веса слова наиболее часто используется TF-IDF (Term Frequency-Inverse Document Frequency) взвешивание. TF-IDF взвешивание базируется на мысли о том, что чем чаще слово встречается во всех документах, тем хуже оно идентифицирует конкретный.

Степень схожести двух документов определяется значением косинуса их векторов.

Процедуру построения рекомендаций для документа можно разбить на 6 шагов:

1) Из документов удаляется вся разметка и стоп-слова (слова, которые не могут быть использованы в качестве смысловых, например, предлоги), оставшиеся слова приводятся в их корням;

2) Для каждого слова в документе вычисляется его вес

$$tf = \frac{(\text{сколько раз слово встречается в текущем документе})}{(\text{число слов в текущем документе})}$$

$$idf = \log \frac{(\text{общее число документов})}{(\text{число документов, в которых хоть раз было это слово})}$$

$$\text{вес слова} = tf * idf$$

3) Считается степень схожести двух документов

$$dp = (\text{вес слова 1 документа 1} * \text{вес слова 1 документа 2} + \text{вес слова 2 документа 1} * \text{вес слова 2 документа 2} + \cdots + \text{вес слова } n \text{ документа 1} * \text{вес слова } n \text{ документа 2})$$

$$d1 = \sqrt{\text{вес слова 1 документа } 1^2 + \text{вес слова 2 документа } 1^2 + \cdots}$$

$$d2 = \sqrt{\text{вес слова 1 документа } 2^2 + \text{вес слова 2 документа } 2^2 + \cdots}$$

$$\text{степень схожести двух документов} = \frac{dp}{d1 * d2}$$

4) Документы, с наибольшей степенью схожести является рекомендуемыми для текущего документа.

Данный алгоритм является универсальным и подходит для всех документов, содержащих текстовую информацию. В настоящее время электронные документы хранятся в специальных форматах, которые позволяют форматировать содержимое и задавать определенную структуру, такую как оглавления, нумерация разделов, таблиц, изображений и т. д. Если добавить учет особенности этой структуры к классическому рекомендательному алгоритму, то можно улучшить качество рекомендаций.

В качестве демонстрации данной идеи, я взял 30 статей с сайта “Хабрахабр” [2], поскольку они содержат большое количество слов, имеют структуру, а самое главное, правила сайта не запрещают использование их материалов в подобных целях. Список статей:

1) Bitcoin. Как это работает [3]
2) Bitcoin как приманка правоохранительных органов или нацбезопасности-основные аргументы [4]
3) Bitcoin — объяснение экспоненциального роста [5]
4) Bitcoin. Стоит ли доверять [6]
5) DDOS любого сайта с помощью Google Spreadsheet [7]
6) DirectX 12 [8]
7) Google Docs индексирует PDF [9]
8) Google ввел шифрование Gmail-трафика между дата-центрами для надежной защиты данных пользователей [10]
9) Qt Bitcoin Trader — программа для торговли Bitcoin под Windows, Mac и Linux [11]
10) Банкомат. Некоторые особенности [12]
11) В Gmail и Google Docs появилось распознавание рукописного текста [13]
12) В Стэнфордском университете разработали бумажный микроскоп стоимостью меньше доллара [14]

13) Генерация Bitcoin в браузере — Обратная сторона [15]
14) Джаббер переходит на полное шифрование [16]
15) Единое зарядное устройство для всех мобильных девайсов — в ЕС приняли новый закон [17]
16) Как инфраструктура Яндекс.Почты выросла за 13 лет [18]
17) Как мы тестируем рекламные технологии Яндекса, и как этому научиться [19]
18) Как статистика помогает делать Яндекс.Пробки лучше [20]
19) Новый интерфейс Яндекс.Метро и технологии, с помощью которых он работает [21]
20) Облачная платформа Яндекса — подробнее про Elliptics [22]
21) Облачная платформа Яндекса. Cocaine [23]
22) Рассекречена личность Сатоси Накамото [24]
23) Релиз KPHP и движков [25]
24) Самодельный фазовый лазерный дальномер [26]
25) Светлое будущее IPv6 — когда уже наконец наступит новый мировой порядок [27]
26) Фундаментальные проблемы экономики на Bitcoin [28]
27) Чемоданы айфонов и гопники в Бутово — как мы чуть не разорились на продукции Apple [29]
28) Яндекс открывает офис разработки в Берлине [30]
29) Яндекс против шокирующей рекламы [31]
30) Яндекс.Кит — новая прошивка для смартфонов [32]

Для примера, я буду увеличивать вес лексем в заголовках разделов и названиях статей, поскольку они фактически, описывают смысл статьи и её абзацев.

Результаты обычного алгоритма, для статьи “Bitcoin. Как это работает” выглядят следующим образом:

Таблица №1 – пример №1 результатов обычного алгоритма

№	Коэффициент схожести	Название статьи
1	0,129260802276918	Bitcoin. Стоит ли доверять
2	0,109382554741675	Bitcoin как приманка правоохранительных органов или нацбезопасности — основные аргументы
3	0,106182491029751	Генерация Bitcoin в браузере — Обратная сторона
4	0,0996296385183458	Bitcoin — объяснение экспоненциального роста
5	0,0634968195044193	Qt Bitcoin Trader — программа для торговли Bitcoin под Windows, Mac и Linux
6	0,0590696045772304	Облачная платформа Яндекса — подробнее про Elliptics
7	0,0546396941545713	Фундаментальные проблемы экономики на Bitcoin

Практически каждая из статей содержит внутри себя заголовки разделов. Например, в статье “Bitcoin. Как это работает” они следующие:

Bitcoin. Как это работает;

Настоящие деньги?;

Сложность добычи;

Условно ограниченный ресурс;

Материальность;

Цепочка блоков;

Блок;

Транзакции;

Заключение;

Точно так же, как и в обычном рекомендательном алгоритме, будем

удалять стоп-слова и символы пунктуации. Далее, получаем коэффициент схожести структур двух документов по следующей формуле:

$$\frac{(\text{число слов в заголовках обоих документов})}{(\text{число слов в заголовках первого документа} + \text{число слов в заголовке второго})}$$

Этот коэффициент будем прибавлять к степени схожести двух документов из первого алгоритма. Он будет равен нулю, если в заголовках разделов нет ни одного общего слова, и поэтому никак не будет изменять результат обычного алгоритма. Если же общие слова имеются, то этот коэффициент немного увеличит коэффициент схожести. В случае, когда совпадают абсолютно все слова в обоих документах, этот коэффциент будет равен 0,5. На мой взгляд, такой случай возможен только лишь у двух копий документов.

Результаты для статьи “Bitcoin. Как это работает” после введения учета схожести заголовков разделов документов:

Таблица №2 – пример №1 результатов изменённого алгоритма

№	Коэффициент схожести	Название статьи
1	0,15602195557258	Генерация Bitcoin в браузере — Обратная сторона
2	0,152598966997783	Bitcoin как приманка правоохранительных органов или нацбезопасности — основные аргументы
3	0,139857847606885	Bitcoin. Стоит ли доверять
4	0,139552367375821	Qt Bitcoin Trader — программа для торговли Bitcoin под Windows, Mac и Linux
5	0,139542115011107	Bitcoin — объяснение экспоненциального роста
6	0,0929130198673841	Фундаментальные проблемы экономики на Bitcoin
7	0,0911585411446258	Облачная платформа Яндекса — подробнее про

		Elliptics

Статья “Облачная платформа Яндекса — подробнее про Elliptics” про key-value хранилище Яндекса, она попала к результатам из-за обилия слов транзакция, нода, группа, ключ, блок и т.д. Обычный алгоритм ранжировал её на 6-ю позицию, а алгоритм, учитывающий заголовки, на 7-ю, уступив место статье “Фундаментальные проблемы экономики на Bitcoin”, которая, безусловно, по смыслу намного ближе к статье, для которой мы искали рекомендации.

Результаты работы алгоритмов, для статьи “Google ввел шифрование Gmail-трафика между дата-центрами для надежной защиты данных пользователей”

Таблица №3 – пример №2 результатов обычного алгоритма

№	Коэффициент схожести	Название статьи
1	0,118333103738983	Как инфраструктура Яндекс.Почты выросла за 13 лет
2	0,107647145272587	В Gmail и Google Docs появилось распознавание рукописного текста
3	0,0962958787016505	DDOS любого сайта с помощью Google Spreadsheet
4	0,0746002645405106	Bitcoin как приманка правоохранительных органов или нацбезопасности- основные аргументы
5	0,0697387994316876	Облачная платформа Яндекса — подробнее про Elliptics
6	0,0667618780052418	Google Docs индексирует PDF
7	0,0658406422133608	Светлое будущее IPv6- когда уже наконец наступит новый мировой порядок

Таблица №4 – пример №2 результатов изменённого алгоритма

№	Коэффициент схожести	Название статьи
1	0,166470674684352	В Gmail и Google Docs появилось распознавание рукописного текста
2	0,15879587870165	DDOS любого сайта с помощью Google Spreadsheet
3	0,138190449433813	Google Docs индексирует PDF
4	0,118333103738983	Как инфраструктура Яндекс.Почты выросла за 13 лет
5	0,0829064883508363	Джаббер переходит на полное шифрование
6	0,0746002645405106	Bitcoin как приманка правоохранительных органов или нацбезопасности- основные аргументы
7	0,0697387994316876	Облачная платформа Яндекса- подробнее про Elliptics

Как видно из данных примеров, отдельный учет слов в заголовках документов позволяет получить более качественную рекомендательную выборку.

Список литературы:

1. Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor – Recommender Systems Handbook – 2010
2. «Хабрахабр» [Электронный ресурс] URL: http://habrahabr.ru/
3. «Bitcoin. Как это работает» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/114642/ (дата обращения: 10.04.2014)
4. «Bitcoin как приманка правоохранительных органов или нацбезопасности — основные аргументы» [Электронный ресурс] //

Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/198006/ (дата обращения: 10.04.2014)

5. «Bitcoin — объяснение экспоненциального роста» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/203530/ (дата обращения: 10.04.2014)
6. «Bitcoin. Стоит ли доверять» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/203518/ (дата обращения: 10.04.2014)
7. «DDOS любого сайта с помощью Google Spreadsheet» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/215233/ (дата обращения: 10.04.2014)
8. «DirectX 12» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/microsoft/blog/216579/ (дата обращения: 10.04.2014)
9. «Google Docs индексирует PDF» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/78526/ (дата обращения: 10.04.2014)
10. «Google ввел шифрование Gmail-трафика между дата-центрами для надежной защиты данных пользователей» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/216557/ (дата обращения: 10.04.2014)
11. «Qt Bitcoin Trader — программа для торговли Bitcoin под Windows, Mac и Linux» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/183536/ (дата обращения: 10.04.2014)
12. «Банкомат. Некоторые особенности» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/216315/ (дата обращения: 10.04.2014)
13. «В Gmail и Google Docs появилось распознавание рукописного текста» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/198860/ (дата обращения: 10.04.2014)

14. «В Стэнфордском университете разработали бумажный микроскоп стоимостью меньше доллара» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/215223/ (дата обращения: 10.04.2014)
15. «Генерация Bitcoin в браузере — Обратная сторона» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/119749/ (дата обращения: 10.04.2014)
16. «Джаббер переходит на полное шифрование» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/208338/ (дата обращения: 10.04.2014)
17. «Единое зарядное устройство для всех мобильных девайсов - в ЕС приняли новый закон» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/216395/ (дата обращения: 10.04.2014)
18. «Как инфраструктура Яндекс.Почты выросла за 13 лет» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/210478/ (дата обращения: 10.04.2014)
19. «Как мы тестируем рекламные технологии Яндекса, и как этому научиться» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/212875/ (дата обращения: 10.04.2014)
20. «Как статистика помогает делать Яндекс.Пробки лучше» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/210240/ (дата обращения: 10.04.2014)
21. «Новый интерфейс Яндекс.Метро и технологии, с помощью которых он работает» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/213033/ (дата обращения: 10.04.2014)

22. «Облачная платформа Яндекса — подробнее про Elliptics» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/209882/ (дата обращения: 10.04.2014)
23. «Облачная платформа Яндекса. Cocaine» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/209324/ (дата обращения: 10.04.2014)
24. «Рассекречена личность Сатоси Накамото» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/214903/ (дата обращения: 10.04.2014)
25. «Релиз KPHP и движков» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/vkontakte/blog/214877/ (дата обращения: 10.04.2014)
26. «Самодельный фазовый лазерный дальномер» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/213749/ (дата обращения: 10.04.2014)
27. «Светлое будущее IPv6 — когда уже наконец наступит новый мировой порядок» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/215535/ (дата обращения: 10.04.2014)
28. «Фундаментальные проблемы экономики на Bitcoin» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/post/181356/ (дата обращения: 10.04.2014)
29. «Чемоданы айфонов и гопники в Бутово - как мы чуть не разорились на продукции Apple» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/madrobots/blog/214529/ (дата обращения: 10.04.2014)

30. «Яндекс открывает офис разработки в Берлине» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/211632/ (дата обращения: 10.04.2014)
31. «Яндекс против шокирующей рекламы» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/210712/ (дата обращения: 10.04.2014)
32. «Яндекс.Кит — новая прошивка для смартфонов» [Электронный ресурс] // Хабрахабр: коллективный блог URL: http://habrahabr.ru/company/yandex/blog/213103/ (дата обращения: 10.04.2014)

дата публикации: 11.04.2014