УДК 004.891
МРНТИ 28.23.29 DOI 10.55452/1998-6688-2021-18-1-109-116

ИНТЕЛЛЕКТУАЛЬНЫЙ МОДУЛЬ ДЛЯ «УМНОГО» НОВОСТНОГО АГРЕГАТОРА

ИБРАГИМ Н.А.

Казахский Национальный университет имени аль-Фараби

Аннотация: *В сегодняшнее время все больше людей получают информацию с онлайн ресурсов, таких как новостные порталы, блоги и т.п. С развитием интернет технологий объем публикуемой информации настолько вырос, что стало трудно и долго находить релевантную и интересную информацию. Новостные агрегаторы – это решение, которое предоставляет возможность пользователю получать только свежие и релевантные новости с разных источников. Платформа агрегатора контента собирает информацию со всей сети и публикует ее в одном месте для доступа посетителей. В данной работе представлена интеллектуальная система новостного агрегатора, которая собирает свежие новости с разных источников с помощью канала RSS/Atom и выводит их в одной платформе. В новостном агрегаторе реализован интеллектуальный модуль, который на основе сохраненных пользователями новостей рекомендует похожие новости. Для рекомендации пользователям похожих новостей к новостным заголовкам применяется метод косинусного сходства, который измеряет схожесть двух векторов путем вычисления косинуса угла между этими двумя векторами. Таким образом, новостные заголовки, которые имеют наибольшее значение косинусного сходства, рекомендуются пользователям. К новостным заголовком применяются следующие технологии обработки естественного языка: токенизация, удаление ненужных символов и пунктуаций, преобразование заголовков в вектора с помощью метода TF-IDF. В данной работе были сравнены результаты измерения сходства для самых популярных метрик, таких как косинусное сходство, Евклидово расстояние и расстояние Жаккарда. Результаты сравнения представлены для новостей, полученных через RSS/Atom каналы ресурсов из категорий программирование и бизнес/маркетинг.*

Ключевые слова: *новостной агрегатор, RSS, Atom, интеллектуальный модуль, предварительная обработка текста, TF-IDF, рекомендация новостей, косинусное сходство*

«АҚЫЛДЫ» ЖАҢАЛЫҚТАР АГРЕГАТОРЫНЫҢ ИНТЕЛЛЕКТУАЛДЫ МОДУЛІ

Аңдатпа: *Қазіргі кезде жаңалық порталдары, блогтар және т.б. сияқты интернет-ресурстардан ақпарат алатындар саны артып келеді. Интернет-технологиялардың дамуына байланысты жарияланған ақпараттың көлемі өскені соншалық, өзекті және қызықты ақпаратты табу қиын әрі ұзаққа созылады. Жаңалықтар агрегаторлары – бұл қолданушыға сан түрлі ақпарат көздерінен тек жаңа және өзекті жаңалықтарды алуға мүмкіндік беретін шешім. Мазмұнды біріктіретін платформа бүкіл ғаламтордан ақпарат жинайды және оны пайдаланушылар қолдану үшін бір жерде жариялайды. Бұл мақалада RSS / Atom арнасын қолдана отырып, әртүрлі ақпарат көздерінен соңғы жаңалықтарды жинайтын және оларды бір платформада көрсететін интеллектуалды жаңалықтар агрегатор жүйесі қарастырылған. Жаңалықтар агрегаторында интеллектуалды модуль іске асырылған, ол пайдаланушылар сақтаған жаңалықтар негізінде оларға ұқсас жаңалықтарды ұсынады. Ұқсас жаңалықтарды қолданушыларға ұсыну үшін жаңалықтар тақырыптарына косинустық ұқсастық әдісі қолданылады, бұл екі вектордың арасындағы косинусты есептеу арқылы екі вектордың ұқсастығын өлшейді. Осылайша косинустық ұқсастық мәні ең жоғары жаңалықтар тақырыптары пайдаланушыларға беріледі. Жаңалықтар тақырыбына табиғи тілді өңдеудің келесі технологиялары қолда-*

нылады: токенизациялау, қажет емес таңбалар мен пунктуацияларды жою, тақырыптарды TF-IDF әдісі арқылы векторларға айналдыру. Бұл жұмыста косинустық ұқсастық, евклидтік арақашықтық және Жаккард арақашықтығы сияқты ең танымал метрикалар үшін ұқсастық өлшемдері салыстырылды. Салыстыру нәтижелері бағдарламалау және бизнес / маркетинг санаттарынан RSS / Atom ресурстық арналары арқылы алынған жаңалықтар үшін ұсынылған.

Түйінді сөздер: *жаңалық агрегаторы, RSS, Atom, интеллектуалды модуль, мәтінді алдын ала өңдеу, TF-IDF, жаңалықтар ұсынысы, косинустық ұқсастық*

INTELLIGENT MODULE FOR «SMART» NEWS AGGREGATOR

Abstract: *Nowadays more and more people get information from online resources such as news portals, blogs, etc. With the development of Internet technologies, the volume of published information has grown so much that it has become difficult and long to find relevant and interesting information. News aggregators are a solution that allows the user to receive only fresh and relevant news from various sources. The content aggregator platform collects information from all over the web and publishes it in one place for visitors to access. This paper presents an intelligent news aggregator system that collects the latest news from different sources using an RSS / Atom feed and displays them in one platform. The news aggregator has an intelligent module that recommends similar news based on the news saved by users. In order to recommend similar news to users, the cosine similarity method is applied to news headlines, which measures the similarity of two vectors by calculating the cosine of the angle between the two vectors. Thus, the news headlines that have the highest cosine similarity value are recommended to users. The following natural language processing technologies are applied to the news headline: tokenization, removing unnecessary characters and punctuation, converting headlines to vectors using the TF-IDF method. In this paper, similarity measurements were compared for the most popular metrics, such as cosine similarity, Euclidean distance, and Jaccard distance. Comparison results are presented for news received via RSS / Atom resource feeds from the programming and business / marketing categories.*

Key words: *news aggregator, RSS, Atom, intelligent module, text preprocessing, TF-IDF, news recommendation, cosine similarity*

Введение

Сегодня Интернет предоставляет нам возможность создавать и взаимодействовать с цифровым контентом на разных источниках и платформах. Опубликованная информация может быть доступна в любой точке мира за несколько кликов [1]. С развитием интернет технологий объем публикуемой информации онлайн ресурсов растет с каждым днем. Растущий объем публикуемой информации привел к тому, что конечным пользователям становится все труднее находить необходимое им информацию. Поэтому за последние несколько лет появились агрегаторы контента, которые собирают информацию из множества информационных каналов. Новостной агрегатор является агрегатором новостной информации, который автоматически собирает публикуемые новости из разных источников и показывает их на одной странице. С помощью новостного агрегатора пользователю не придется переходить по разным вкладкам браузера для просмотра публикуемых новостей. Все нужные источники будут находиться на одной странице, что экономит время просмотра новостной информации. Как и текстовые форматы, мультимедийный контент и полноформатное видео также распространяются с помощью агрегаторов контента [2]. К таким агрегаторам контента можно отнести программное обеспечение ITunes для музыкального контента и Youtube, который предоставляет мультимедийный контент.

Также популярными видами агрегаторов контента являются агрегаторы социальных сетей. Агрегаторы социальных сетей собирают информации из разных социальных сетей и блогов и объединяют их в один источник [3]. Главными функциями агрегаторов социальных сетей являются совместное использование, голосование и обсуждение постов [4]. Одними из популярных агрегаторов социальных сетей являются Reddit и Hacker News [5].

Новостной агрегатор, представленный в текущей работе, автоматически собирает новости из других источников с помощью каналов синдикации RSS и Atom. RSS и Atom являются одними из известных и популярных форматов каналов для синдикации контента. Они основаны на файловом формате XML и предоставляют стандартизированный способ публикации и подписки на контент [6]. В текущей работе пользователь имеет возможность подписаться на ресурсы, которые предоставляют свои новости через RSS канал. После подписки новостной агрегатор будет получать опубликованные новости подписанных ресурсов через этот канал и будет выводить их на своей странице. Объединение разных источников в одной платформе дает пользователю возможность просматривать новости на одной странице.

Особенностью новостного агрегатора, представленного в данной работе, является внедрение интеллектуального модуля. Интеллектуальный модуль основан на алгоритмах машинного обучения, которые позволяют анализировать новостные заголовки и извлекать из них полезные данные. Интеллектуальная составляющая новостного агрегатора автоматически рекомендует новости пользователю на основе сохраненных им новостей.

1 Архитектура разработанного новостного агрегатора

Новостной агрегатор, представленный в текущей работе, собирает новости с источников с помощью RSS и Atom каналов. Новости ресурсов, которые предоставляют данные каналы, будут отображаться в платформе. При нажатии на новость пользователь переходит на страницу самого ресурса для дальнейшего прочтения новости.

Платформа собирает ресурсы, которые предоставляют данные каналы.

Общие требования к новостному агрегатору

Операционная система: Unix-системы, Windows

Браузеры: Chrome, Opera, Firefox и т.д.

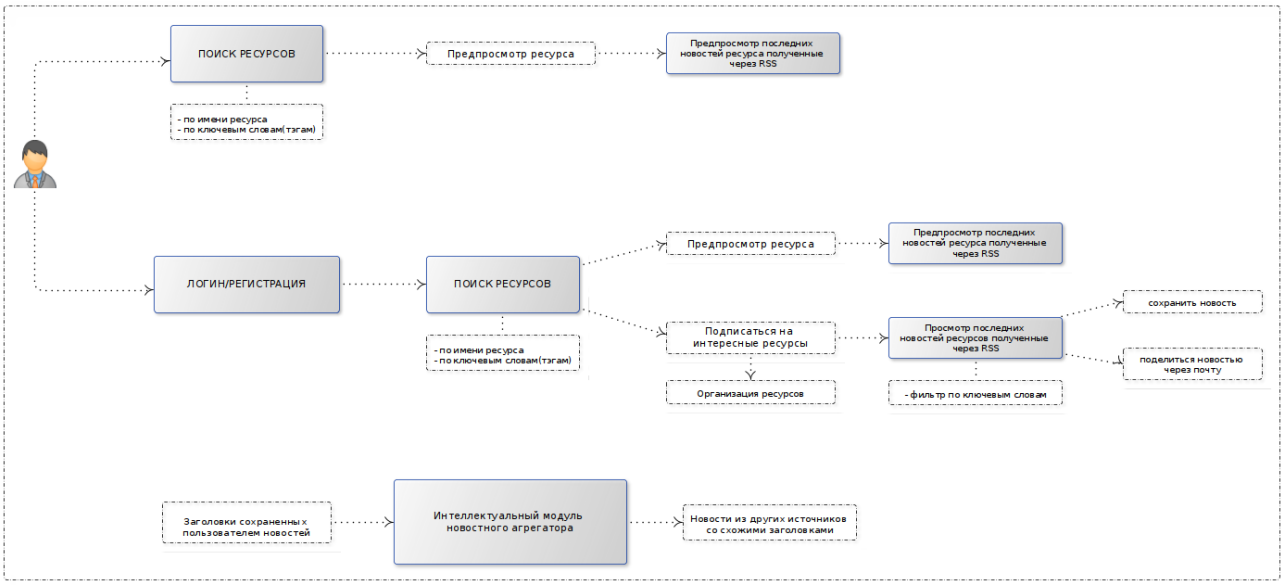

Рис. 1 – Архитектура новостного агрегатора

Используемые технологии

Разработка: Python, веб-фреймворк Django, Celery

База данных: PostgreSQL

Кэширование: Redis

2 Интеллектуальный модуль

В данной работе разработан модуль для новостного агрегатора.

Интеллектуальный модуль анализирует заголовки новостей и рекомендует пользователям новости со схожими заголовками используя технологии обработки естественного языка(NLP – Natural Language Processing). Рекомендательная система является эффективным методом, которая дает возможность пользователю находить интересные и нужные информации без самостоятельного поиска среди несколько разных источников.

Интеллектуальный модуль новостного агрегатора на вход принимает сохраненные пользователем заголовки новостей и на выход выдает самые похожие заголовки с разных источников. Задача данного модуля состоит в работе с заголовками новостей, а именно анализ и вытаскивание нужной информации с текста. Алгоритмы машинного обучения не могут работать напрямую с необработанным текстом, поэтому перед анализом текст необходимо преобразовать в векторы чисел.

В общем виде работа интеллектуального модуля состоит из следующих процессов

- предварительная обработка заголовков
- извлечение признаков
- измерение косинусного сходства заголовков
- сортировка и выдача самых похожих заголовков

2.1 Предварительная обработка заголовков

Первый шаг в обработке естественного языка – это предварительная обработка. Предварительной обработкой текста называется процесс подготовки текста перед процессом кодирования. В работах [7, 8] описываются методы предобработки текста. Эти методы состоят из токенизации, удалении стоп-слов, приведении всех букв в нижний регистр и стемминга. В этих работах описывается влияние предварительной обработки при классификации текстов. В работе говорится, что влияние стемминга и удалении стоп-слов имеют разное влияние на точность алгоритма классификации для разных исследований. Предлагается использовать удаление стоп-слов и стемминг для уменьшения размерности признаков и увеличения эффективности алгоритмов машинного обучения. Для алгоритма стемминга в текущей работе выбран алгоритм Стеммера Портера. Алгоритм Стеммера Портера часто применяется для стемминга английского языка [7].

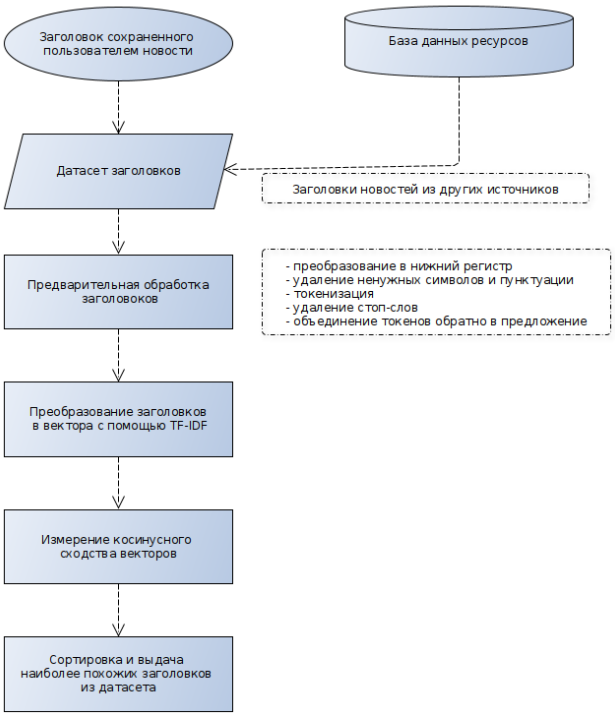

Рис. 2 – Алгоритм работы интеллектуального модуля

2.2 Извлечение признаков

Так как компьютеры не понимают текст на естественном языке, их необходимо преобразовать в векторы чисел.

При языковой обработке векторы x выводятся из текстовых данных, чтобы отразить различные лингвистические свойства текста [9]. В данной статье [9] описаны простые и популярные методы извлечения признаков из текста такие как Bag-Of-Words, TF-IDF и Word2Vec. TFIDF, сокращенно от слова «ча-

стота документа с обратной частотой», представляет собой числовую статистику, которая предназначена для отражения того, насколько важно слово для документа в коллекции или корпусе. Значение TF-IDF увеличивается пропорционально количеству раз слово появляется в документе и компенсируется количеством документов в корпусе, которые содержат это слово, что помогает учесть тот факт, что некоторые слова в целом встречаются чаще. TF-IDF – одна из самых популярных сегодня методом взвешивания терминов; 83% текстовых рекомендательных систем в электронных библиотеках используют TF-IDF [9]. В интеллектуальном модуле, который представлен в данной работе для извлечения признаков с заголовков был выбран метод TF-IDF.

$$TF - IDF = TF(i,j) \times IDF(i)$$

В результате применения метода TF-IDF формируется числовой вектор, который представляет в закодированном виде извлеченные заголовки новостей. Данное представление в дальнейшем облегчает процесс сравнения заголовков. Данное сравнение может быть осуществлено с помощью кластерных алгоритмов [10]. В настоящей работе применяется косинусное сходство, описанное в пункте 3.3.

Косинусное сходство

Косинусное сходство измеряет схожесть двух векторов путем вычисления косинуса угла между этими двумя векторами. Для измерения меры сходства векторов обычно используется скалярное произведение. В работе[11] проведено исследование применения косинусного сходства для вычисления сходства входных и выходных векторов. Косинусное сходство – одна из наиболее широко используемых и мощных мер подобия в Data Science. Он используется во многих приложениях, таких как поиск похожих документов в NLP, поиск информации, поиск аналогичной последовательности ДНК в биоинформатике, обнаружение плагиата и многое другое [12].

Методы, основанные на подобии, определяют наиболее похожие объекты с наивысшими значениями, поскольку это подразумевает, что они живут в более близких окрестностях. Из этих методов для анализа текста самым популярным является косинусное сходство. Для вычисления косинусного сходство используется следующая формула:

$$cos\,(\theta) = \frac{A \times B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Косинусное сходство в первую очередь касается ориентации, а не величины. Чем меньше угол между векторами, тем более похожими они являются. Поэтому в случаях когда величина векторов не важна, например при измерении дистанций между векторами, используется косинусное сходство [13]. В текущей работе косинусное сходство применяется для нахождения похожих заголовков новостей.

3 Эксперимент

3.1 Подготовка датасета

Для проведения эксперимента были выбраны источники из 2 категорий: программирование и бизнес/маркетинг. Из категории программирование было выбрано 7 ресурсов, такие как Real Python, Clean Coder Blog и т.п. Для категории бизнес/маркетинг было отобрано 4 ресурса: Entrepreneur, Bloomberg и т.п. Датасет был построен на основе заголовков новостей, которые были получены через RSS/Atom каналы всех выбранных ресурсов. В итоге было получено 140 новостных заголовков для категории программирование и 105 для категории бизнес/маркетинг

3.2 Сравнение метрик

Для сравнения метрик были выбраны следующие три метрики измерения дистанций между векторами: косинусное сходство, Евклидово расстояние и расстояние Жаккарда.

Для проверки работы интеллектуального модуля на вход для категории программирование был подан заголовок “Cool New Features in Python 3.9”. Также для категории бизнес/маркетинг был выбран заголо-

Таблица 1 – Результат выдачи похожих заголовков 3-х метрик для категории программирование

Метод	Ресурс	Заголовок	Величина
Косинусное сходство	Real Python	Python 3.9: Cool New Features for You to Try	0.863150096550615
	Real Python	The Real Python Podcast – Episode #30: Exploring the New Features of Python 3.9	0.5593503873802708
Евклидово расстояние	Real Python	Python 3.9: Cool New Features for You to Try	0.5231632698295726
	Real Python	The Real Python Podcast – Episode #30: Exploring the New Features of Python 3.9	0.938775385936092
Расстояние Жаккарда	Real Python	The Real Python Podcast – Episode #30: Exploring the New Features of Python 3.9	0.782608695652174
	Real Python	The Real Python Podcast – Episode #33: Going Beyond the Basic Stuff With Python and Al Sweigart	0.7727272727272727

Таблица 2 – Результат выдачи похожих заголовков 3-х метрик для категории бизнес/ маркетинг

Метод	Ресурс	Заголовок	Величина
Косинусное сходство	Tech	Robocalls, WeChat messages, and more spread misinformation on Election Day	0.23818083754550962
	Bloomberg Politics	Two Bad Election Scenarios Come Back to Haunt Global Markets	0.20250327408707736
Евклидово расстояние	Tech	Robocalls, WeChat messages, and more spread misinformation on Election Day	1.2343574542688116
	Bloomberg Politics	Two Bad Election Scenarios Come Back to Haunt Global Markets	1.2629305015818746
Расстояние Жаккарда	Quartz	Nigeria's EndSARS protests have been about much more than police brutality	0.8947368421052632
	Quartz	What's the ethical case for CEOs publicly endorsing candidates?	0.8947368421052632

вок “Some scenarios of what could happen on election day and beyond”.

Заключение

В данной статье был представлен новостной агрегатор, который автоматически собирает новости из различных источников с помощью RSS/Atom каналы. Особенностью новостного агрегатора, представленного в данной работе, является внедрение интеллектуального модуля для рекомендации похожих новостей. В работе [14] представлен новый подход для рекомендации научных статей на основе объединения мультимодальных представлений. Рекомендательная система представленная в работе [15] использовала документы слушателей для извлечения контекстной рекомендации. В интеллектуальном модуле, который представлен в текущей работе, рекомендация новостей осуществляется с помощью преобразования заголовков в вектора методом TF-IDF и измерения косинусного сходства полученных векторов. Заголовки, которые имеют наибольшее значение косинусного сходства будут рекомендоваться пользователям. В дальнейших работах планируется внедрение в новостной агрегатор возможность регистрации с помощью Google аккаунта и аккаунтов социальных сетей. Зарегистрированные пользователи будут получать уведомления на почту о публикации новых новостей с подписанных ресурсов.

ЛИТЕРАТУРА

1. Sudatta Chowdhury Monica Landoni. "News aggregator services: user expectations and experience" // Online Information Review.– 2006. – Т 30. –100-115 с.
2. William A. Hanff. News aggregator [Электронный ресурс].-URL: https://www.britannica.com/topic/news-aggregator
3. Агрегатор социальных сетей: материал из Википедии [Электронный ресурс].-URL: https://en.wikipedia.org/wiki/News_aggregator
4. Franziska Zimmer. An Evaluation of the Social News Aggregator Reddit // European Conference on Social Media. – 2018. – Лимерик, Ирландия.
5. Adrienne Erin. 10 social news aggregators to help you reach new audiences [Электронный ресурс].-URL: https://socialnomics.net/2015/01/08/10-social-news-aggregators-to-help-you-reach-new-audiences/
6. Alex Stolz, Martin Hepp. From RDF to RSS and Atom: Content Syndication with Linked Data // 24th ACM Conference on Hypertext and Social Media. – 1-3 Мая 2013. – Париж, Франция.
7. V. Srividhya, R. Anitha. Evaluating Preprocessing Techniques in Text Categorization // International Journal of Computer Science and Application Issue.-2010.
8. Dr. S. Vijayarani, MS. J. Ilamathi, Ms. Nithya. Preprocessing Techniques for Text Mining - An Overview // International Journalof Computer Science & Communication Networks. – Т 5(1). – 7-16 с.
9. Prasoon Singh. Fundamentals of Bag Of Words and TF-IDF [Электронный ресурс].-URL: https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22
10. Korbinian Koch. A friendly introduction to text clustering [Электронный ресурс].-URL: https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04
11. Tan Thongtan, Tanasanee Phienthrakul. Sentiment Classification using Document Embeddings trained with Cosine Similarity // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop.-28 Июля-2 Августа 2019. – Флоренция, Италия. – 407-414 с.
12. Varun. Cosine similarity: How does it measure the similarity, Maths behind and usage in Python [Электронный ресурс].-URL: https://towardsdatascience.com/cosine-similarity-how-does-it-measure-the-similarity-maths-behind-and-usage-in-python-50ad30aad7db

13. Chris Emmery. Euclidean vs. Cosine Distance [Электронный ресурс].-URL: https://cmry.github.io/notes/euclidean-v-cosine#:~:text=Cosine%20similarity%20is%20generally%20used,data%20represented%20by%20word%20counts.
14. Shashank Gupta, Vasudeva Varma. Scientific Article Recommendation by using Distributed Representations of Text and Graph // International World Wide Web Conference Committee (IW3C2). – 2017.
15. Ziwon Hyung, Kibeom Lee, Kyogu Lee. Music recommendation using text analysis on song requests to radio stations // Music and Audio Research Group, Graduate School of Convergence Science and Technology, Seoul National University. – 2013. – Сеул, Корея.