Рекомендательная система классификации физико-математических документов

Ш.М. Хайдаров, Г.Ш. Ямалутдинова

Институт математики и механики им. Н.И. Лобачевского Казанского (Приволжского) федерального университета

Аннотация. Обсуждены проблемы возрастания значения научных классификаторов для систематизации научной информации в цифровую эпоху, так, например, классификация документов (присвоение индексов-классификаторов) является традиционным способом систематизации знаний и поиска информации. В настоящей работе предложена рекомендательная система автоматизированного подбора индексов Универсальной десятичной классификации (УДК) для физико-математических документов. Эта система реализует один из сервисов цифровой математической библиотеки Lobachevskii-DML. Предложенный алгоритм основан на использовании терминов, извлеченных из названия, списка ключевых слов и аннотации, приведенных в анализируемых документах. Извлечение терминов из документов коллекции проводится с помощью разработанных нами программных инструментов, учитывающих стилевые особенности оформления документов и положения в тексте искомых терминов. Полученные данные были включены в словарь, который имеет структуру инвертированного индекса. Сформированный словарь содержит как классификационные признаки, так и наборы ключевых терминов, по которым производятся систематизация и классификация материала. Большая часть этих терминов была получена путем автоматизированной обработки коллекции архивов физико-математических публикаций Общероссийского математического портала Math-Net.Ru. Предложен вариант семантической разметки таблицы индексов универсальной десятичной классификации. Описана модель классификации научных документов.

Ключевые слова: рекомендательная система, систематизация научной информации, классификаторы научной информации, системы классификации, УДК, извлечение информации

Recommender System of Physical and Mathematical Documents Classification

S.M. Khaydarov, G.S. Yamalutdinova

N.I. Lobachevskii Institute of Mathematics and Mechanics
Kazan (Volga Region) Federal University

Abstract. The problems of increasing the value of scientific classifiers for the systematization of scientific information in the digital age are discussed, for example, the classification of documents (assignment of indices-classifiers) is a traditional way of systematization of knowledge and information search. In this paper we propose a recommendation system for automated selection of Universal decimal classification (UDC) indices for physical and mathematical documents. This system implements one of the services of the digital mathematical library Lobachevskii-DML. The proposed algorithm is based on the use of terms extracted from the title, the list of keywords and annotations given in the analyzed documents. Extraction of terms from the documents of the collection is carried out with the help of software tools developed by us, taking into account the stylistic features of the documents and the positions in the text of the required terms. The data obtained were included in a dictionary that has an inverted index structure. The generated dictionary contains both classification features and sets of key terms, which are used to systematize and classify the material. Most of these terms were obtained by automated processing of the collection of archives of physical and mathematical publications of the All-Russian mathematical portal Math-Net.Ru. The proposed variant of the semantic markup of a table of indices of the Universal decimal classification. The model of classification of scientific documents is described.

Keywords: recommender system, systematization of scientific information, scientific information classifiers, classification system, UDC, information extraction

Введение

В современной научной деятельности одной из актуальных задач информационного поиска является классификация документов, заключающаяся в их отнесении к одной или нескольким категориям на основании содержания документов. Для этого разработаны различные системы классификации, такие, например, как Библиотечно-библиографическая классификация (ББК) [1], Государственный рубрикатор научно-технической информации (ГРНТИ) [2], Mathematics Subject Classification (MSC2010) [3]. Одной из наиболее широко распространенных систем является Универсальная десятичная классификация (УДК) [4].

Классификация документов (присвоение индексов-классификаторов) является традиционным способом систематизации знаний и поиска информации. В научных документах классификаторы служат одним из видов метаданных (см., например, [5]). Подбор классификационных индексов сопряжен с анали-

зом структуры дерева классификаторов и достаточно трудоемок. Для упрощения этого процесса необходимо решить задачу автоматизации такого подбора.

Как известно (см., например, [6, 7]), рекомендательными (рекомендующими) системами называют класс систем принятия решений, которые используют знания об интересах и предпочтениях человека для прогнозирования его реакции воспользоваться некоторой услугой. Выделяют два основных типа рекомендательных систем: контент-ориентированные и социальные (коллаборативной фильтрации). Первые основаны на представлении предпочтений пользователей путем анализа содержимого рекомендательных элементов, а системы второго типа моделируют предпочтения, оценивая близость профилей пользователей. Отметим, что в рекомендательных системах для физико-математического контента используются также соответствующие предметные онтологии [8, 9].

В настоящей работе предложена рекомендательная система автоматизированного подбора классификаторов УДК для физико-математических документов. Эта система реализует один из сервисов цифровой математической библиотеки Lobachevskii-DML [10].

1. Структура рекомендательной системы

Алгоритм функционирования созданной рекомендательной системы представлен на Рис. 1. Базовой составной частью этой системы является словарь классификационных терминов, описанный в следующем разделе. Предложенный алгоритм основан на использовании терминов, извлеченных из названия, списка ключевых слов и аннотации, приведенных в документах.

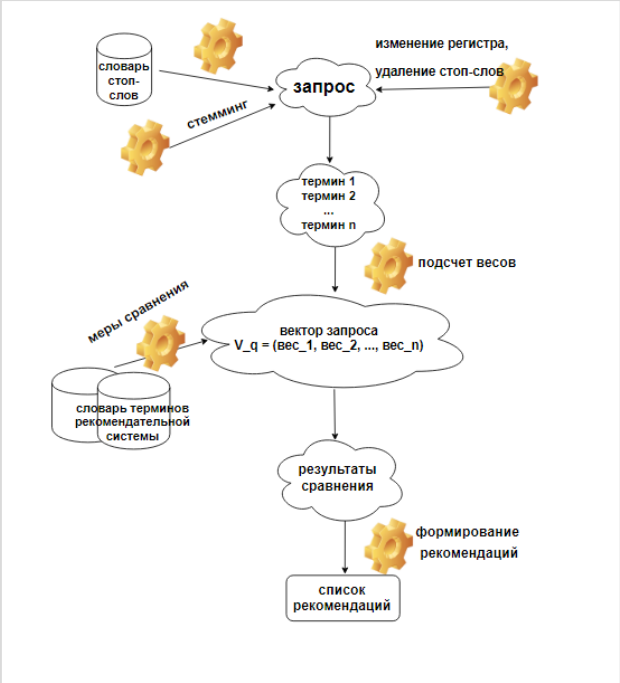

Рис. 1. Алгоритм функционирования рекомендательной системы

Словарь терминов, ассоциированных с классификаторами УДК

Сформированный словарь содержит как классификационные признаки, так и наборы ключевых терминов, по которым производятся систематизация и классификация материала. Большая часть этих терминов была получена путем автоматизированной обработки коллекции архивов физико-математических публикаций Общероссийского математического портала Math-Net.Ru [11]. На этом портале размещено более 200 тысяч статей, в 57 тысячах из которых указан индекс УДК. Мы исходили из предположения, что в этих публикациях все классификационные индексы указаны авторами правильно. Для пополнения словаря эти статьи были обработаны, в результате чего были выделены метаданные (название, ключевые слова и аннотация), из которых были извлечены термины. Полученные данные были включены в словарь (см. [12]), который имеет структуру инвертированного индекса в виде семантического представления и является частью создаваемой рекомендательной системы.

Как известно (см., например, [13, 14]), распространенной моделью информационного поиска является модель векторного пространства. В этой модели документы представляются в виде векторов $\bar{V}(d_i)$, компонентами которых являются веса терминов, где под документами мы подразумеваем индексы-классификаторы. Существуют различные схемы взвешивания терминов, одной из которых является статическая мера

$$tfidf(t,d,D) = tf(t,d)\ idf(t,D),$$

где D – множество документов, d – документ, t – термин. Известно несколько модификаций функций $tf(t,d)$ и $idf(t,D)$. Например,

$$tf(t,d) = \frac{n_{t,d}}{n_d}, \qquad idf(t,D) = \log\left(\frac{|D|}{|\{d_i \in D | t \in d_i\}|}\right),$$

где $n_{t,d}$ – количество вхождений термина t в документ d, n_d – общее количество слов в документе d, $|D|$ – количество документов в множестве, $|\{d_i \in D | t \in d_i\}|$ – количество документов, в которых встречается термин t (см. [15]);

$$tfidf(t,d,D) = \left(\sum_{t \in d} \begin{matrix} 1\ if\ t \in d \\ 0\ else \end{matrix}\right) \log\left(\frac{N-n}{n}\right),$$

где N – количество документов в множестве, n – количество документов, в которых встречается термин t (см. [16]);

$$tfidf(t,d,D) = tf(t,d)\ \max\left\{0,\ \log\left(\frac{N-DF(t)}{DF(t)}\right)\right\},$$

где $DF(t)$ – число документов в коллекции, где встречается термин (см. [17]).

Для оценки сходства между документами, представленными в виде векторов, мы применяем косинусную меру

$$sim(d_1,d_2) = \frac{\left(\bar{V}(d_1), \bar{V}(d_2)\right)}{\|\bar{V}(d_1)\|\|\bar{V}(d_2)\|},$$

а также функцию ранжирования Okapi BM25

$$\text{score}(\text{d}, \text{Q}) = \sum_{j=1}^{n} \text{idf}(t_j) \frac{f(t_j,d)(k_1+1)}{f(t_j,d)+k_1\left(1-b+b\left(\frac{|d|}{avgdl}\right)\right)}.$$

Для сокращения временных затрат составлен инвертированный индекс, в котором каждому термину поставлен в соответствие номер документа, содержащего этот термин.

Работа выполнена за счет средств субсидии, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности, проект 1.2368.2017/ПЧ, и при частичной финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта №18-47-160012.

Литература

1. Библиотечно-библиографическая классификация. – URL: http://roslavl.library67.ru/files/382/bbk.pdf.
2. Государственный рубрикатор научно-технической информации. – URL: http://grnti.ru/.
3. Классификатор математических сущностей MSC2010. – URL: http://www.ams.org/mathscinet/msc/msc2010.html.
4. UDC Summary Linked Data. – URL: http://udcdata.info/.
5. Елизаров А.М., Зуев Д.С., Липачёв Е.К. Управление жизненным циклом электронных публикаций в информационной системе научного журнала // Вестник Воронеж. гос. ун-та. Сер. Систем. анализ и информ. технологии, 2014. – № 4. – С. 81–88.
6. Ricci F., Rokach L., Shapira B., Kantor P.B. (Eds.) Recommender Systems Handbook. Springer-Verlag New York. 2011. 842 p. https://doi.org/10.1007/978-0-387-85820-3.
7. Ricci F., Rokach L., and Shapira B. (Eds.) Recommender Systems Handbook. – Springer-Verlag New York, 2015. 1003 p. https://doi.org/10.1007/978-1-4899-7637-6.
8. Елизаров А.М., Липачёв Е.К., Невзорова О.А., Соловьев В.Д. Методы и средства семантического структурирования электронных математических документов // Доклады РАН. – 2014. – Т. 457 (6). – С. 642–645. https://doi.org/10.7868/S0869565214240049.
9. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов // Доклады РАН. – 2016. – Т. 467, № 4. – С. 392–395. https://doi.org/10.7868/S0869565216100042.
10. Елизаров А.М., Липачёв Е.К. Семантические методы и инструменты электронной математической библиотеки Lobachevskii-DML // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18–23 сентября 2017 г., г. Новороссийск). – М.: ИПМ им. М.В. Келдыша,

2017. – С. 130–136. – URL: http://keldysh.ru/abrau/2017/73.pdf. https://doi.org/ 10.20948/abrau-2017-73.
11. Общероссийский математический портал Math-Net.Ru. – URL: http://www.mathnet.ru/.
12. Хайдаров Ш.М., Ямалутдинова Г.Ш. Алгоритм формирования словарей рекомендующей системы подбора классификаторов научной информации // Ученые записки ИСГЗ. – 2017. – Т. 15. – №1. – С. 552–557.
13. Ингерсолл Г.С., Мортон Т.С., Фэррис Э.Л. Обработка неструктурированных текстов. Поиск, организация и манипулирование. / Пер. с англ. Слинкин А.А. – М.: ДМК Пресс, 2015. – 414 с.
14. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск: Пер. с англ. – М.: ООО «И. Д. Вильямс», 2014. – 528 с.
15. Недильченко О.С. Этапы и методы автоматического извлечения ключевых слов // Молодой ученый. – 2017. – №22 (156). – С. 60–62.
16. Резников И.А. Обзор алгоритмов извлечения ключевых слов из текста // 58-я научная конференция Московского физико-технического института (23–28 ноября 2015 г., г. Долгопрудный). – URL: http://conf58.mipt.ru/static/ reports_pdf/499.pdf.
17. Нокель М.А., Лукашевич Н.В. Тематические модели в извлечении однословных терминов // Программная инженерия. – Издательство «Новые технологии» (Москва), 2015. – №3. – С. 34–40.

References

1. Bibliotechno-bibliograficheskaja klassifikacija. – URL: http://roslavl.library67.ru/ files/382/bbk.pdf.
2. Gosudarstvennyj rubrikator nauchno-tehnicheskoj informacii. – URL: http://grnti.ru/.
3. 2010 Mathematics Subject Classification. – URL: http://www.ams.org/ mathscinet/msc/msc2010.html.
4. UDC Summary Linked Data. – URL: http://udcdata.info/.
5. Elizarov A.M., Zuev D.S., and Lipachev E.K. Upravlenie zhiznennym ciklom jelektronnyh publikacij v informacionnoj sisteme nauchnogo zhurnala // Vestnik Voronezh. gos. un-ta. Ser. Sistem. analiz i inform. tehnologii, 2014. – No. 4. – P. 81–88.
6. Ricci F., Rokach L., Shapira B., and Kantor P.B. (Eds.) Recommender Systems Handbook. Springer-Verlag New York. 2011. 842 p. https://doi.org/10.1007/978-0-387-85820-3.
7. Ricci F., Rokach L., and Shapira B. (Eds.) Recommender Systems Handbook. – Springer-Verlag New York, 2015. 1003 p. https://doi.org/10.1007/978-1-4899-7637-6.
8. Elizarov A.M., Lipachev E.K., Nevzorova O.A., and Solov’ev V.D. Methods and means for semantic structuring of electronic mathematical documents //

Doklady Mathematics. – 2014. – Vol. 90 (1). – P. 521–524. https://doi.org/10.1134/ S1064562414050275.
9. Elizarov A.M., Kirillovich A.V., Lipachev E.K., Zhizhchenko A.B., and Zhil'tsov N.G. Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics // Doklady Mathematics. – 2016. – Vol. 93 (2). – P. 231–233. https://doi.org/10.1134/S1064562416020174.
10. Elizarov A.M., and Lipachev E.K. Semanticheskie metody i instrumenty jelektronnoj matematicheskoj biblioteki Lobachevskii-DML // Nauchnyj servis v seti Internet: trudy XIX Vserossijskoj nauchnoj konferencii (18–23 September 2017, g. Novorossiysk). – M.: IPM im. M.V. Keldysha, 2017. – S. 130–136. – URL: http://keldysh.ru/abrau/2017/73.pdf. https://doi.org/10.20948/abrau-2017-73.
11. All-Russian Mathematical Portal Math-Net.Ru. – URL: http://www.mathnet.ru/.
12. Khaydarov S.M., and Yamalutdinova G.S. Algorithm for forming the dictionary of the recommender system of selecting classifiers of scientific information // Uchenye zapiski ISGZ. – 2017. – Vol. 15 (1). – P. 552–557.
13. Ingersoll G.S., Morton T.S., and Farris D. Taming Text: How to Find, Organize, and Manipulate It – M.: DMK Press, 2015. – 414 p.
14. Manning C.D., Raghavan P., and Schütze H. Introduction to Information Retrieval – M.: OOO «I. D. Vil'jams», 2014. – 528 p.
15. Nedil'chenko O.S. Jetapy i metody avtomaticheskogo izvlechenija kljuchevyh slov // Molodoj uchenyj. – 2017. – No. 22 (156). – P. 60–62.
16. Reznikov I.A. Obzor algoritmov izvlechenija kljuchevyh slov iz teksta // 58-ja nauchnaja konferencija Moskovskogo fiziko-tehnicheskogo instituta (23–28 November 2015, Dolgoprudny). – URL: http://conf58.mipt.ru/ static/reports_pdf/ 499.pdf.
17. Nokel' M.A., and Lukashevich N.V. Tematicheskie modeli v izvlechenii od-noslovnyh terminov // Programmnaja inzhenerija. – Izdatel'stvo «Novye tehnologii» (Moscow), 2015. – No. 3. – P. 34–40.