Терминологическое аннотирование и рекомендательный сервис в системе управления физико-математическим контентом

© А. М. Елизаров © Н. Г. Жильцов © А. В. Кириллович © Е. К. Липачёв
Казанский (Приволжский) федеральный университет
Казань
amelizarov nikita.zhiltsov alik.kirillovich elipachev {@gmail.com}

Аннотация

Рассмотрены методы формирования семантического профиля пользователя информационных систем управления научным контентом. Онтологические модели предметных областей и методы терминологического аннотирования использованы в качестве технологической основы построения рекомендательного сервиса, обеспечивающего персонализированный отбор научных документов, релевантных семантическому профилю учёного.

1 Введение

В настоящее время объем информации, размещенной в Сети, настолько велик, что пользователь уже не может его обработать, применяя традиционные инструменты и подходы. Это стало значительной проблемой, одним из способов решения которой стала разработка рекомендательных сервисов. По определению, рекомендательный сервис должен отбирать из всего объема имеющейся информации ту, которая соответствует интересам пользователя. Такие сервисы широко распространены, например, в интернет-магазинах, социальных сетях и специализированных электронных библиотеках, в частности, музыкальных. В связи с ростом объемов научной информации несомненна полезность рекомендательных сервисов, связанных с научными публикациями. Они решают следующую задачу: подобрать пользователю, заинтересовавшемуся некоторой статьей, другие статьи, которые также могут его заинтересовать. Такие сервисы позволяют выделить «похожие» публикации – таковыми считают публикации, близкие по ряду выделенных признаков (тематическое подобие, различные меры близости и др., см., например, [1, 2]).

Рекомендательными сервисами обладают известные системы работы с научными публикациями:

- поисковая система Google Scholar [3], которая для заданной статьи находит «похожие» статьи (кнопка «Похожие статьи» на странице публикации) и составляет для пользователя индивидуальный список рекомендованных публикаций (раздел «Мои обновления») на основе его цитат;
- реферативная база данных Scopus [4], которая на странице каждой публикации отображает список похожих публикаций (блок «Related documents» на странице публикации), сервис основан на похожих авторах и ключевых словах;
- система управления библиографической информацией Mendeley [5], которая на странице публикации отображает список похожих публикаций (блок «Related Full-Text Papers» на этой странице);
- электронная библиотека eLIBRARY.ru [6], которая для заданной статьи находит похожие статьи (ссылка «Найти близкие по тематике публикации» на странице публикации).

Традиционно список похожих публикаций формируется на основе близости ссылок и составляемого авторами списка ключевых слов. Вместе с тем, использование ключевых слов не лишено рядом недостатков, в частности:

- список ключевых слов может быть неполным или вовсе отсутствовать;
- одно и то же понятие может обозначаться разными ключевыми словами (проблема омонимии), например, «полином» и «многочлен»;
- не учитываются родо-видовые отношения между понятиями, например, статья с ключевым словом «матрица» не будет похожей на статью с ключевым словом «лямбда-матрица»;
- существенна привязка к языку, например, статья с ключевым словом «матрица» не будет похожа на статью с ключевым словом «matrix».

Таким образом, авторского списка ключевых слов недостаточно – необходимо глубже анализировать содержание материала. Одним из методов такого анализа является терминологическое аннотирование, основанное на онтологиях предметных областей (например, [7 – 10]). Один из примеров терми-

нологического аннотирования – в системе ArXiv [11] на основе онтологии ScienceWise [12], но только для английского языка; в ней, собственно, рекомендации не формируются.

Задача, решаемая нами, связана с созданием рекомендательного сервиса, способного обрабатывать русскоязычные публикации. В качестве коллекции публикаций используется база данных Math-Net.Ru [13].

Цели настоящей работы – автоматизация процессов поиска публикаций по теме исследований, а также получение дополнительной информации о терминологии, используемой в анализируемых публикациях.

2 Пользовательский интерфейс рекомендательного сервиса

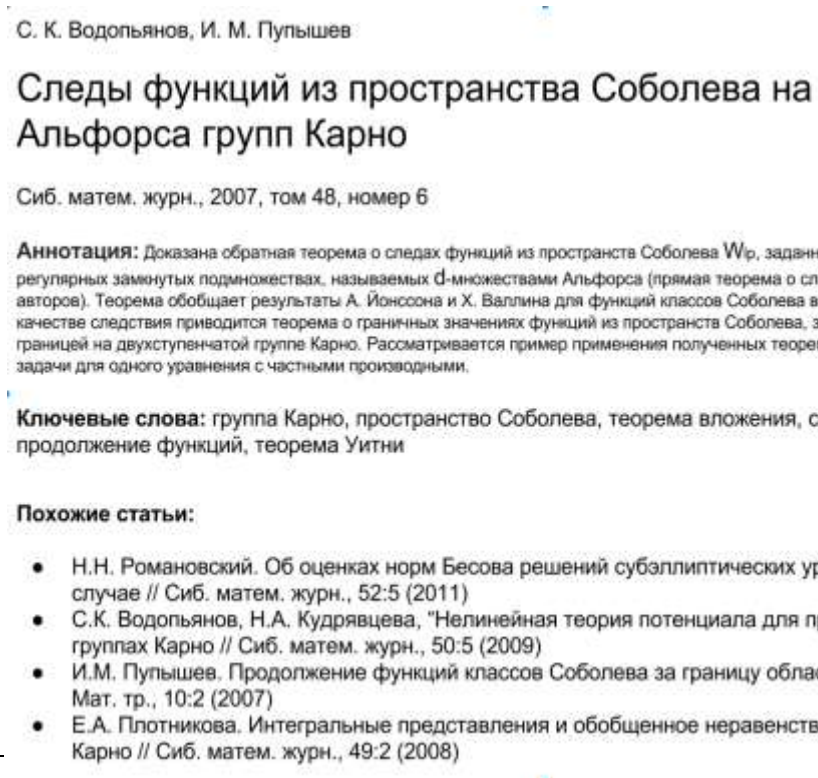
Рекомендательный сервис разработан как надстройка портала Math-Net.Ru, содержащего информацию о математических публикациях, журналах, авторах, исследовательских организациях и т. д., и расширяет карточки публикаций, имеющихся на портале, и формирует карточки терминов.

Карточка публикации содержит метаданные научной публикации: заголовок, имя автора, журнал, авторский список ключевых слов и т. д. Список ключевых слов присутствует не всегда. Сервис добавляет два дополнительных блока:

- автоматически построенный список ключевых слов (гиперссылки, которые указывают на карточку термина);
- список похожих публикаций (гиперссылки, которые указывают на карточку соответствующей публикации).

На рис. 1 приведен один из примеров работы сервиса с контентом портала Math-Net.Ru. Блоки ключевых слов и похожих статей сформированы автоматически. Каждое ключевое слово – гиперссылка, ведущая на карточку термина (см. рис. 2).
Карточка термина содержит информацию о заданном математическом понятии, его положении в понятийном пространстве и генерируется автоматически на основе онтологии OntoMathPro [14]. Карточка содержит:

- название термина;
- более общие и более частные термины (отношения «is–a» онтологии OntoMathPro) – гиперссылки, которые указывают на карточки соответствующих терминов;
- определение термина из OntoMathPro;
- ссылки на информацию о термине на внешних ресурсах: OntoMathPRO, MSC 2010, Wolfram Mathworld, The Wolfram Functions Site, ScienceWISE, Математическая энциклопедия, Википедия;
- список публикаций, посвященных термину, – гиперссылки, указывающие на карточки публикаций.



С. К. Водопьянов, И. М. Пупышев

Следы функций из пространства Соболева на
Альфорса групп Карно

Сиб. матем. журн., 2007, том 48, номер 6

Аннотация: Доказана обратная теорема о следах функций из пространств Соболева W_p, заданн
регулярных замкнутых подмножествах, называемых d-множествами Альфорса (прямая теорема о сл
авторов). Теорема обобщает результаты А. Йонссона и Х. Валлина для функций классов Соболева в
качестве следствия приводится теорема о граничных значениях функций из пространств Соболева, з
границей на двухступенчатой группе Карно. Рассматривается пример применения полученных теоре
задачи для одного уравнения с частными производными.

Ключевые слова: группа Карно, пространство Соболева, теорема вложения, с
продолжение функций, теорема Уитни

Похожие статьи:

- Н.Н. Романовский. Об оценках норм Бесова решений субэллиптических ур
случае // Сиб. матем. журн., 52:5 (2011)
- С.К. Водопьянов, Н.А. Кудрявцева, "Нелинейная теория потенциала для п
группах Карно // Сиб. матем. журн., 50:5 (2009)
- И.М. Пупышев. Продолжение функций классов Соболева за границу обла
Мат. тр., 10:2 (2007)
- Е.А. Плотникова. Интегральные представления и обобщенное неравенств
Карно // Сиб. матем. журн., 49:2 (2008)

Рис. 1. Расширенная карточка публикации

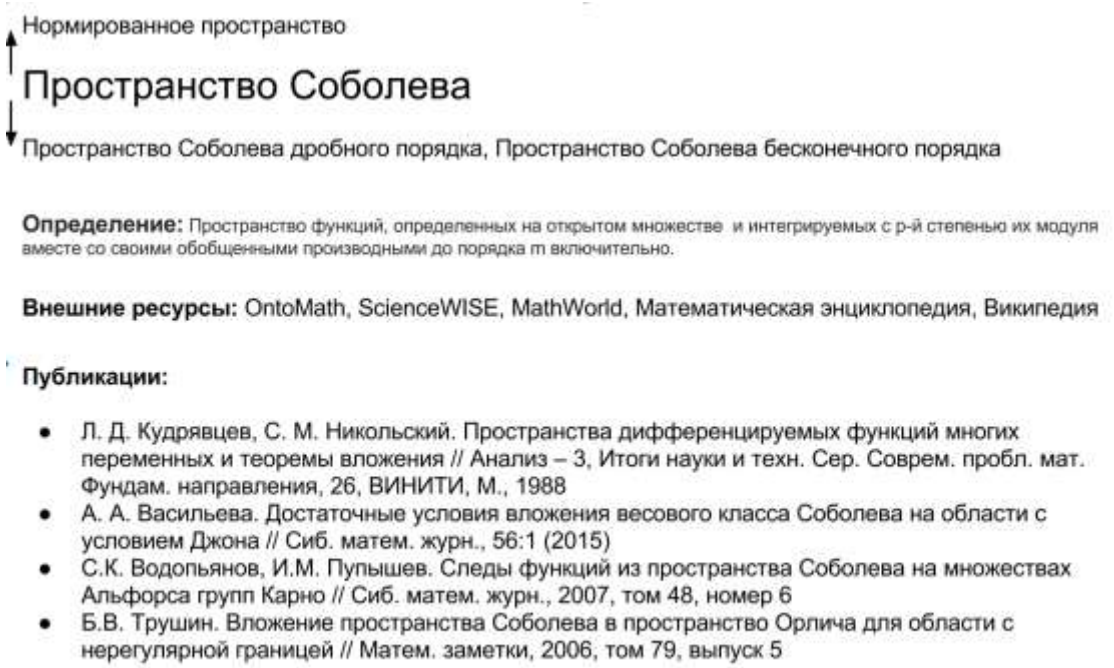


Рис. 2. Карточка термина

Центральным элементом системы является онтология OntoMath^{PRO} (http://ontomathpro.org/), разработанная при участии авторов [14, 15]. Она содержит описания математических понятий и связей между ними.

Понятия содержат:

- заголовок: русский и английский;
- определение
- ссылки на внешние ресурсы из наборов DBpedia [16] и ScienceWISE [12];
- связи с другими концептами.

Типы связей:

- Класс → Подкласс (Число → Простое число)
- Область математики → Математический объект (Метрическая геометрия → Барицентрические координаты)
- Определяется с помощью (Символ Кристоффеля → Связность)
- Смотри также (Циклический итерационный метод Чебышева → Численное решение СЛУ)
- Задача → Метод решения (Система линейных уравнений → Метод Гаусса)

Модуль извлечения ключевых слов принимает на вход текст и метаданные статьи и возвращает список извлеченных ключевых слов с весами. Модуль работает в три этапа (подробнее см. [15]).

На первом этапе происходит извлечение кандидатов в ключевые слова. В качестве кандидатов рассматриваются все именные группы. Именные группы извлекаются с помощью облачной платформы текстовой аналитики Textocat API, разработанной авторами [17].

На втором этапе из кандидатов в ключевые слова отбираются те, которые обозначают математические понятия. В основе метода лежит нахождение степени соответствия между кандидатом и именем понятия.

Степень соответствия между кандидатом и именем понятия онтологии представляет собой число из диапазона $[0; 1]$ и определяется правилами:

- если имя понятия не содержит главное слово кандидата, то степень соответствия = 0.
- если имя понятия длиннее (по числу токенов), чем кандидат, то степень соответствия = 0.
- в противном случае степень соответствия вычисляется как коэффициент Жаккара на множествах токенов имени понятия и термина (множественная интерпретация).

Кандидаты с мерой близости больше пороговой рассматриваются как ключевые слова.

На третьем этапе ключевым словам присваивается вес. Вес вычисляется на основе меры TF-IDF и положения термина в логической структуре публикации (термин в заголовке весит больше, чем в формулировке теоремы, а в формулировке весит выше, чем доказательстве).

Извлеченные слова сохраняются в базе данных сервиса и отображаются на карточке публикации.

Модуль поиска похожих публикаций строит список похожих публикаций. Публикации представляются в виде вектора, каждый элемент которого является весом соответствующего понятия из онтологии OntoMathPro для данной публикации (вычисляется в предыдущем модуле). Между публикациями вычисляется мера близости на основе косинусной меры близости их векторов. Рекомендуемые публикации — публикации с мерой близости, меньшей заданного порога.

Заключение

Представлена информационная система, автоматизирующая процесс поиска публикаций по теме исследований, а также получение дополнительной информации о терминологии, используемой в анализируемых публикациях. В качестве технологической основы построения рекомендательного сервиса, обеспечивающего персонализированный отбор научных документов, релевантных семантическому профилю учёного, использованы онтологические модели предметных областей и методы терминологического аннотирования.

Работа выполнена при финансовой поддержке РФФИ (проекты 15-07-08522, 15-47-02472).

Литература

[1] Елизаров А.М., Липачев Е.К., Малахальцев М.А. Веб-технологии для математика: Основы MathML. Практическое руководство. – М.: Физматлит, 2010. – 216 с.

[2] Шарнин М.М., Петров А.В., Кузнецов И.П. Методика учета интересов пользователя при работе в сети Internet на основе его профиля и ассоциативных связей // Труды 15-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии» – RCDL'2013, г. Ярославль, 14-17 октября 2013 г. – С. 86-90.

[3] Захаров В.Н., Хорошилов А.А. Автоматическое формирование визуального представления смыслового содержания // Системы и средства информации. – 2013. – Т. 23, Вып. 1. – С. 143-158.

[4] Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьев В.Д. Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Сер. 2. Информ. процессы и системы. – 2014. – №4. – С. 12-17.

[5] Захаров В.Н., Хорошилов А.А., Хорошилов А.А. Опыт создания кластеров документов на основе метода определения их тематического подобия // Труды 16-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии» – RCDL'2014, г. Дубна, 13-16 октября 2014 г. – С. 322-328.

[6] Ding L., Kolari P., Ding Z., Avancha S. Using ontologies in the Semantic Web: a survey// Ontologies, Springer US. – 2007. – P. 79-113.

[7] Ермаков А.Е. Автоматизация онтологического инжиниринга в системах извлечения знаний из текста // Материалы межд. науч. конф. «Диалог 2008», М., 2008. – С. 154-159.

[8] Норенков И.П. Интеллектуальные технологии на основе онтологий // Информационные технологии. – 2010. – № 1. – С. 17-23.

[9] Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G. Mathematical knowledge representation: se-mantic models and formalisms // Lobachevskii J. of Mathematics. – 2014. – V. 35, No 4. – P. 347-353.

[10] Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Иванов В.В., Невзорова О.А., Соловьев В.Д. Модель семантического поиска в коллекциях математических документов на основе онтологий // Труды 12-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии» – RCDL'2010, г. Казань, 13-17 октября 2010 г. – С. 296-300.

[11] Aberer K., Boyarsky A., Cudr-Mauroux P., Demartini G., Ruchayskiy O. ScienceWISE: A Web-based interactive semantic platform for scientific collaboration // 10th Int. Semantic Web Conference (ISWC 2011 – Demo), 2011.

[12] Textocat API: облачный сервис текстовой аналитики. – URL: http://textocat.com.

[13] Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E. Bringing Math to LOD: a semantic publishing platform prototype for scientific collections in mathematics//12th Int. Semantic Web Conference, Sydney, NSW, Australia, October 21–25, 2013. Proceedings, Part I. 8218. Springer Berlin Heidelberg, 2013. – P. 379-394.

[14] Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E. OntoMathPRO ontology: a linked data hub for mathematics // In Knowledge Engineering and the Semantic Web. Springer Int. Publishing. – 2014. – P. 105-119.

[15] Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z., Dbpedia: a nucleus for a Web of open data // In the Semantic Web. – Springer Berlin Heidelberg, 2007. – P. 722-735.

Methods of ontological modeling of natural science knowledge areas

Alexander M. Elizarov, Nikita G. Zhiltsov,
Alexander V. Kirillovich, Evgeny K. Lipachev

Methods of formation of semantic user profile management information systems scientific content presented. Ontological model subject areas and methods of terminology used in the annotation process as a basis for building an advisory service that provides personalized selection of scientific documents relevant semantic profile scientist.