ИНФОРМАТИКА

УДК 004.912

ОНТОЛОГИИ МАТЕМАТИЧЕСКОГО ЗНАНИЯ И РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА ДЛЯ КОЛЛЕКЦИЙ ФИЗИКО-МАТЕМАТИЧЕСКИХ ДОКУМЕНТОВ

А.М. Елизаров[1], А.Б. Жижченко[2], Н.Г. Жильцов[1], А.В. Кириллович[1], Е.К. Липачёв[1]

Одним из способов повышения эффективности работы с контентом в Вебе является использование сервисов, предоставляемых различными системами подготовки рекомендаций. Такие сервисы внедрены в популярные онлайн-библиотеки: поисковую систему Google Scholar, реферативную базу данных Scopus, систему управления библиографической информацией Mendeley, электронную библиотеку eLIBRARY.ru и другие. Существует два основных типа рекомендательных систем: контент-ориентированные и социальные (коллаборативной фильтрации) (см., например, [1] – [4]). Первые основаны на представлении предпочтений пользователей путем анализа содержимого рекомендательных элементов. Системы второго типа моделируют предпочтения, оценивая близость профилей пользователей. Ниже под рекомендательной системой будем понимать информационную систему, которая:

- формирует концептуальное представление электронной коллекции (например, MathNet) на основе онтологий предметных областей;
- вычисляет меру тематической близости документов, используя это представление;
- выдает список документов, отражающих информационные интересы пользователя.

По сравнению с информационно-поисковыми системами рекомендательные системы наиболее полезны, когда у пользователя возникают трудности с формулировкой эффективного поискового запроса. Чаще всего такие трудности возникают при работе с научным контентом.

Для подготовки качественных рекомендаций наличие профиля пользователя, построенного только на основе истории просмотра им документов анализируемой коллекции, недостаточно — необходимо учитывать модели предметных областей, в частности, с помощью онтологий, а также пользовательские сценарии. Пользователями рекомендательной системы могут быть:

- **автор публикации**, которому при работе с коллекцией документов наиболее интересны близкие постановки и методы решения поставленной задачи;

[1]Казанский (Приволжский) федеральный университет

[2]Межведомственный суперкомпьютерный центр РАН

- **рецензент**, который оценивает новизну и актуальность конкретной научного документа по сравнению с другими имеющимися документами по заданной теме;
- **ученый**, который ищет материалы по тематике своих исследования, а также **читатель**, которого интересуют документы, помогающие ему войти в тематику исследований и объясняющие необходимые базовые понятия.

Например, при анализе текста одной и той же статьи автору предполагаемой новой публикации этот материал будет полезен для подготовки обзора работ, близких по тематике, рецензенту будут интересны материалы с определениями специфических терминов, а также близкие статьи для оценки новизны контента, а аспиранту — классические статьи с определениями базовых понятий и основополагающими утверждениями.

В настоящее время онтологии получили распространение как инструмент представления знаний о предметной области, в том числе, для рекомендательных систем [5], благодаря использованию семантических связей. Последние обеспечивают, в частности, повышение полноты и точности поиска, а также мультиязычность. Вместе с тем, одним из препятствий распространению онтологических подходов в рекомендательных системах является трудоемкость создания моделей для каждой предметной области отдельно. Применение онтологического подхода в рекомендательных системах для физико-математического контента стало возможно после появления онтологий физико-математических знаний [6] – [10]. Ниже описана рекомендательная система, основанная на онтологиях математических знаний OntoMath^{PRO} и Mocassin и апробированная на уникальной коллекции физико-математических публикаций MathNet [11, 12].

Онтологии математических знаний. Mocassin – онтология логической структуры математических документов, разработанная нами для автоматического анализа математических публикаций в формате LaTeX. Эта онтология формально (на языке OWL) описывает семантику структурных элементов математических документов (например, теоремы, леммы, доказательства, определения и т. д.), выраженную в виде классов и свойств. Кроме того, онтология содержит аксиомы кардинальности и транзитивности.

OntoMath^{PRO} [6] – [8] содержит около 3500 концептов, которые организованы в две иерархии: математических объектов и областей математики. В этой онтологии определены антисимметричные отношения:

- “Подкласс” → “Класс” (“Многообразие” → “Множество”);
- “Определяется с помощью” (“Атлас” → “Многообразие”),

а также три типа симметричных отношений:

- “Ассоциативная связь” (“Дифференцируемое многообразие” → “Поток”);
- “Задача” → “Метод решения” (“Система линейных уравнений” → “Метод Гаусса”);
- “Область математики” → “Математический объект” (“Алгебраическая топология” → “Группа гомологий”).

Опишем *модель использования структуры онтологии для представления документов.* Онтологию можно представить в виде ориентированного взвешенного графа $G = \langle T, E \rangle$, вершинами которого являются элементы множества концептов T, а ребрами – элементы множество отношений E. Антисимметричному отношению соответствует одно направленное ребро, а симметричному – два разнонаправленных. Вес ребра w_r задается в зависимости от типа отношения r и сценария работы с системой.

Введем меру близости терминов:

$$A[s,d] = \begin{cases} 1, & s = d, \\ 0, & \text{не существует пути между s и d}, \\ 1/dist(s,d) & \text{в остальных случаях}, \end{cases}$$

где $dist(s,d)$ – длина кратчайшего пути между s и d.

Предлагаемый метод работает с коллекцией математических документов и для каждого документа коллекции формирует список близких статей. Метод состоит из следующих этапов. На первом из каждого документа извлекаются ключевые слова – упоминания терминов, описанных в онтологии OntoMath^{PRO}, и их обозначения в виде элементов математической нотации. На втором этапе анализируется логическая структура документа и семантически размечаются его фрагменты на основе онтологии Mocassin [9]. На третьем этапе строится векторное представление документа, учитывающее терминологический состав, положение терминов в логической структуре и связи терминов в онтологическом графе OntoMath^{PRO}. Далее вычисляется мера близости построенных векторов документов коллекции и формируются рекомендации.

Извлечение логической структуры. На данном этапе происходит извлечение логической структуры документа на основе онтологии Mocassin. Фрагменты документа аннотируются как экземпляры классов онтологии. Информация о принадлежности фрагмента документа определенному классу в дальнейшем используется для взвешивания общих терминов с целью вычисления меры близости документов. Каждому концепту онтологии Mocassin ставится в соответствие вес, определяющий его значимость в структуре документа (наибольшее значение присваивается заголовку).

Извлечение ключевых слов. На этом этапе происходит извлечение математических терминов из текста документа. Особенностью является то, что упоминания терминов

часто встречаются в виде символьных обозначений, элементов математической нотации. В качестве базы терминов используется онтология OntoMath^{PRO}; в качестве инструмента анализа – Textocat API (облачная платформа текстовой аналитики, созданная при участи авторов). Извлечение терминов основано на стандартных подходах разрешения лексической многозначности именованных сущностей [13] и состоит из двух основных этапов: извлечение именных групп в качестве фраз-кандидатов на связывание с терминами онтологии; применение бинарного классификатора, принимающего решение о связывании фраз-кандидатов с терминами онтологии с указанием значения меры уверенности. Затем извлеченные понятия привязываются к переменным, которые обозначают экземпляры данного понятия в математических формулах [14]. В дальнейшем все упоминания этой переменной рассматриваются как упоминания соответствующего термина.

В результате мы получаем множество упоминаний M_p терминов онтологии в документе p, в котором каждый элемент представляет собой кортеж $m = \langle t, l \rangle$, где t – термин онтологии, l – позиция термина в документе.

Построение вектора публикации. На этом этапе формируется векторное представление документа в терминологическом пространстве. В качестве терминологической базы используются термины из онтологии OntoMath^{PRO}. Введем порядок на множестве T.

Документ p представляется в виде вектора, каждая компонента которого выражает вес соответствующего термина в данном документе:

$$v(p) = \left(weight(t_1), weight(t_2), \ldots, weight(t_n)\right), n = |T|,$$

где $weight(t)$ – вес термина t в документе. Каждое упоминание увеличивает вес соответствующего термина; увеличение веса зависит от положения термина в логической структуре документа. Кроме того, упоминание термина увеличивает вес всех связанных с ним терминов в соответствии с мерой близости. Таким образом, каждый компонент вектора вычисляется по формуле

$$weight(t^*) = idf(t^*) \cdot \sum_{\langle t,l \rangle \in M_p} \beta_l \cdot A\left[t, t^*\right],$$

где $idf(t) = \log \dfrac{N}{N_t}$, N – число документов в коллекции, N_t – количество документов, содержащих термин t, а β_l – вес термина, определяемый положением l этого термина в логической структуре документа. Использование idf позволяет уменьшить влияние общеупотребительных терминов (таких, например, как “число” или “метод”), слабо характеризующих специфику отдельного документа.

Рекомендательная система на основе вектора документа формирует набор данных, содержащий список близких документов, имеющихся в коллекции, расширенный перечень ключевых слов, отобранных согласно их весам в векторе публикации. Поскольку каждое ключевое слово соответствует термину онтологии, система формирует рекомендации, содержащие ссылки на различные определения термина, место термина в иерархии онтологии, а также список документов коллекции, содержащих данный термин.

Формирование списка близких статей. На этом этапе для каждого документа коллекции строится набор близких документов. В качестве меры близости документов используется известная косинусная мера близости их векторов. Рекомендуемые документы – это документы с мерой близости, большей установленного порога.

Обоснование сформированной рекомендации. Объяснение списка рекомендаций представляет собой список меток, отражающих наиболее важные, с точки зрения модели, атрибуты $weight(t^*)$, учтенные при вычислении меры близости. Примерами меток являются “близкие термины”, “похожая задача”.

Таким образом, созданная рекомендательная система характеризуется следующими особенностями:

- учитывает профессиональный профиль конкретного пользователя, а также других пользователей, интересующихся данной тематикой;
- формирует различные рекомендации для разных сценариев работы с системой (рецензент; пользователь, входящий в тематику, и т. д.);
- назначаем разные веса разным концептам, например: для научного обзора более важны концепты, обозначающие области математики, чем те, которые характеризуют математические объекты; для начинающего исследовател важны обзорные статьи, содержащие понятия из разных областей математики и достаточно ссылок на первоисточники;
- предполагает дальнейшую интеграцию как с MathNet, так и с другими научными коллекциями;
- допускает пополнение онтологии новыми концептами, включая понятия из “Математической энциклопедии”.

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

Список литературы

[1] *Ricci F., Rokach L., Shapira B., Kantor P.B. (Eds.)* Recommender Systems Handbook. Springer-Verlag New York. 2010.

[2] *Bobadilla J., Ortega F., Hernando A., Gutierrez A.* Recommender systems survey. Knowledge-Based Systems. 2013. V. 46. P. 109-–132.

[3] *Lampropoulos A.S., Tsihrintzis G.A.* Machine Learning Paradigms. Applications in Recommender Systems. Springer International Publishing Switzerland, 2015. 125 p.

[4] *Verbert K., Manouselis N., Ochoa X., Wolpers M., Drachsler H., Bosnic I., Duval E.* Context-aware recommender systems for learning: a survey and future challenges // IEEE Transactions on Learning Technologies. 2012. V. 5. No 4. P. 318–335.

[5] *Middleton S.E., De Roure D., Shadbolt N.R.* Ontology-Based Recommender Systems // In Staab S., Studer R. (Eds.) Handbook on Ontologies. Springer-Verlag. Berlin. Heidelberg. 2009. P. 779–796.

[6] *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.

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

[8] *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMath^{PRO} ontology: a linked data hub for mathematics //In Knowledge Engineering and the Semantic Web. Springer International Publishing. Communications in Computer and Information Science. V. 468. P. 105-119, 2014.

[9] *Solovyev V., Zhiltsov N.* // Proc. of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

[10] *Aberer K., Boyarsky A., Cudré-Mauroux P., Demartini G., Ruchayskiy O.* ScienceWISE: a Web-based Interactive Semantic Platform for scientific collaboration // 10th International Semantic Web Conference (ISWC 2011-Demo), Bonn, 2011.

[11] *Жижченко А.Б., Изаак А.Д.* Информационная система Math-Net.Ru. Применение современных технологий в научной работе математика // УМН. 2007. Т. 62. № 5. С. 107–132.

[12] *Chebukov D., Izaak A., Misurina O., Pupyrev Yu., Zhizhchenko A.* Math-Net.Ru as a digital archive of the Russian mathematical knowledge from the XIX century to today. Lecture Notes in Computer Science. 2013. **7961**. P. 344–348.

[13] *Shen W., Wang J., Han J.* Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions // IEEE Transactions on Knowledge and Data Engineering. 2015. V. 27. Issue 2. P. 443–460.

[14] *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. // In The Semantic Web–ISWC. Springer Berlin Heidelberg. 2013. P. 379-394.