УДК 004.891

Семантический поиск в системе поддержки работы с научными публикациями

В.В. Костин

Вычислительный Центр им. А.А. Дородницына РАН

Система поддержки работы с научными публикациями

В [1], [2] рассмотрены такие элементы системы поддержки работы с научными публикациями (СПРНП) как онтология предметной области, структура системы и её возможности. Система предоставляет возможность обращаться к текстам научных публикаций, аннотировать их, хранить и обмениваться с другими пользователями. Система предоставляет возможность вести совместную научную и образовательную деятельность, обмениваться информацией друг с другом, комментировать черновые работы других членов коллектива, хранить важные с точки зрения текущей задачи научные работы в собственной библиотеке. Пользователи могут изменять, уточнять и дополнять знания, хранящиеся в системе, задавать смысловые связи между сущностями, как уже определённые системой, так и новые, системой ещё не описанные. В случае добавления нового типа связей, данные изменения должны верифицироваться специалистами до добавления в систему.

Онтология пользователя

При работе пользователя в системе формируется онтология его знаний и интересов, имеется возможность выделить период работы над конкретной задачей для определения приоритетов пользователя с точки зрения текущей цели. В онтологии хранится информацию о том, какие труды пользователь сохранил в свою библиотеку, с какими из них он наиболее интенсивно работал – цитировал в своих черновиках, аннотировал в системе, ссылался в обсуждениях с другими пользователями системы. Каждый поисковый запрос пользователя указывает на его интересы и фиксируется в виде метаданных в онтологии, поисковая выборка хранится в системе для того, чтобы у пользователя при необходимости была возможность вернуться к результатам поиска. У пользователя существует возможность вручную изменять свою онтологию, наиболее чётко формирую свои цели, текущие знания и научные интересы.

У пользователя есть возможность задавать отдельные информационные срезы онтологии, соответствующие конкретной задаче. Их суть заключается в следующем: каждый срез онтологии интересов пользователя отвечает конкретной задаче – получению знаний в новой научной области, подборка обучающих материалов для студента, поиск материалов для работы над научной статьёй. При выборе отдельного среза вся работа пользователя в системе отображается только в этом срезе: поиск информации, аннотирование, формирование области знаний, интересов и предпочтений. Соответственно, при поиске материала для текущего исследования может быть использован один срез, а для параллельного обучения студента – другой.

Онтология интересов и знаний пользователя пополняется во время двух процессов. Во-первых, во время просмотра научных публикаций, их аннотирования, добавления в личную библиотеку и взаимодействия и обмена знаниями с другими пользователями

системы, их совместной работы. Во-вторых, в процессе серверного анализа информации, хранящейся в базе знаний и интересов пользователя. В этом случае используется семантический анализ интересов и знаний пользователя, и его результаты предлагаются пользователю в качестве рекомендуемой информации.

Семантическая составляющая поиска

В системе применяется математическая модель предпочтений пользователей коллаборативной рекомендательной системы по оценке товаров[3] для случая электронной библиотеки/совокупности публикаций. В этом методе у каждого пользователя для каждых двух публикаций t_i, t_j ($i \neq j$, $i = 1, \ldots, n$, $j = 1, \ldots, n$) проводится попарное сравнение приоритетов пользователя a_k. В качестве параметра сравнения используются несколько критериев – предпочтение при выборе из результата поиска, предпочтение пользователя при выборе из рекомендуемых материалов, интенсивность работы с материалом – его просмотр, добавление в библиотеку, аннотирование, цитирование, рекомендации другим пользователям. Например, $a_k \in A = \{<, >, \sim=\}$, в этом случае $k = 1, 2, 3$. Таким образом, для N публикаций формируется матрица предпочтений R элементов rij, принимающих значения в соответствии с результатами парных сравнений a_k из множества $\{1, 2, \ldots, k, \ldots\}$. При сравнении интересов пользователей сравниваются их матрицы предпочтения и определяется близость интересов пользователей. Но основе близости пользователей и их приоритетов формируется оценка семантической близости трудов для конкретного пользователя.

Пусть D_n – множество научных публикаций системы, U_m – множество пользователей, а W_k – множество лексических единиц (терминов, словосочетаний). P^{NxN} – матрица предпочтений, p_{ij} – связь между публикациями в данном контексте (срезе), $p_{ij} \in \{-1, 0 \ldots n\}$, где -1 – не связанные друг с другом работы в текущем контексте, 0 – работы заслуживающие одновременного рассмотрения, число задаёт недетерминированное последовательность работ (или групп работ).

При поиске пользователь $Q = \{w\}$, где w – набор лексических конструкций, выделенных из текстового запроса. $Q_1 = Q \cap TZ$, где TZ – тезаурус текущей научной области. $R = Q \setminus Q_1$ – набор исключённых из поиска терминов (редактируется пользователем). При поиске пользователь видит множество Q_1 и R, может их отредактировать. В ответ на поисковый запрос пользователь получается набор документов q. Все просмотренные пользователем документы – v. Документы, которые пользователь добавляет в свою библиотеку, считаются релевантными. Они формируют множество a. В соответствии с действиями пользователя формируется его матрица предпочтений. Пусть $i \in a$, $j \in v$. В этом случае для каждой i-й строки в матрице проставляются соответствующие значения.

ЛИТЕРАТУРА

1. Костин В. В. Обзор семантических моделей, описывающих научные публикации и научно-исследовательскую деятельность. // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

2. Костин В. В. К вопросу создания поддержки работы с научными публикациями // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 4. С. 32–37.

3. Понизовкин Д. М., Амелькин С. А. Математическая модель коллаборативных процессов принятия решений //Программные системы: теория и приложения. – 2011. – Т. 2. – №. 4. – С. 95-99.