

В. А. Васенин, С. А. Афонин, А. С. Козицын, Д. Д. Голомазов, А. В. Бахтин, Г. М. Ганкин (Москва, НИИ механики МГУ). **Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА).**

Доклад посвящен разработанной системе управления информацией, характеризующей деятельность научных и образовательных учреждений. Под управлением информацией понимается ее поиск, учет и систематизация, хранение и тематический анализ. Тематический анализ означает способность на основе анализа данных из коллекции исходных текстов отвечать на различные аналитические запросы. Результатами выполнения таких запросов могут быть: перечень направлений, которые активно исследуются в рамках интересующей области знания (в запросе направление исследования может задаваться набором ключевых слов); динамика интереса исследователей к отдельному направлению исследований во времени; перечень задач, в которых используются методы интересующего направления; список публикаций по конкретному направлению за определенный период времени; список публикаций, похожих на заданную и другие, аналогичные им.

Для решения поставленной задачи управления научной информацией авторами разрабатывается программный комплекс ИСТИНА — Интеллектуальная Система Тематического Исследования Научно-технической информации (см. <http://istina.imec.msu.ru>). Основными задачами системы ИСТИНА являются, во-первых, предоставление возможности сотрудникам структурных подразделений научного учреждения вести учет результатов своей научной деятельности и в автоматизированном режиме формировать годовые научные отчеты, во-вторых, предоставление руководителям отдельных структурных подразделений и организации в целом автоматизированного средства проведения количественного и тематического анализа научной деятельности каждого из сотрудников, отдельных подразделений и всего учреждения.

Для выполнения тематического анализа в системе ИСТИНА используются два типа онтологий. Во-первых, это онтология научной деятельности, которая является расширением онтологии Semantic Web for Research Communities (SWRC, см. <http://istina.imec.msu.ru>) [1]. Она включает в себя такие концепты, как человек, организация, публикация, конференция, проект, а также связи между ними. Ко второму типу онтологий принадлежат модели конкретных областей знания на языке OWL, например, онтология математики, информатики или физики. Для эффективного формирования таких онтологий разработан алгоритм *Sonmake* автоматизированного построения онтологий областей научного знания на основе текстов анонсов научных конференций и информации из поисковых систем в Интернет. Заметим, что в нем активно используется алгоритм выделения терминов *Brainsterm* [2], разработанный одним из авторов статьи.

Построенные онтологии используются, в том числе, для уточнения и расширения запросов при осуществлении пользователем полнотекстового поиска в Интернет.

При уточнении запросов используется информация об онтологии интересов пользователя. Набор заданных пользователем ключевых слов дополняется понятиями онтологии, соответствующих его интересам. Например, запрос «определение группы» будет преобразован в «определение группы AND (математика OR физика OR вычислительные методы)». Такое уточнение позволит системе не показывать пользователю тексты, содержащие определения групп, но имеющих психологическую или биологическую направленность. При расширении запроса используется общая онтология. Указанное в запросе понятие дополняется списком связанных с ним дочерних понятий с учетом типа, степени и направления связи. Например, запрос «алгоритм RSA» будет дополнен понятиями «теорема Эйлера», «открытый ключ» и другими. Такое дополнение позволит найти документы, которые посвящены обсуждению связанных с алгоритмом RSA вопросов, однако не содержат его явного упоминания.

В настоящее время система ИСТИНА позволяет в удобном для конечного пользователя режиме вводить данные о публикациях путем автоматизированного разбора информации из библиографических ссылок и BibTeX-записей. На основе введенных в хранилище системы данных о публикациях для каждого сотрудника автоматически создается отдельная «домашняя» страница, содержащая список его публикаций. В настоящее время проводится внедрение системы ИСТИНА в НИИ механики МГУ имени М. В. Ломоносова, в дальнейшем планируется использовать систему в других подразделениях Московского университета.

Работа выполнена при финансовой поддержке гос. контракта 07.514.11.4116 Министерства образования и науки РФ.

СПИСОК ЛИТЕРАТУРЫ

1. *Sure Y., Blochdorn S., Haase P., Hartmann J., Oberle D.* The swrc ontology — semantic web for research communities. — In: Proceedings of the 12th Portuguese Conference on artificial Intelligence «Progress in Artificial Intelligence» (EPIA 2005). Covilha: Springer, 2005, v. 3803 of LNCS, p. 218–231.
2. *Голомазов Д. Д.* Выделение терминов из коллекции текстов с заданным тематическим делением. — Информационные технологии, 2010, № 2, с. 8–11.