OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 519.711.74

О ПОНЯТИИ ФОРМАЛЬНОЙ КОМПЕТЕНТНОСТИ НАУЧНЫХ СОТРУДНИКОВ

Крюков К.В., Кузнецов О.П., Суховеров В.С.

Институт проблем управления им. В.А. Трапезникова РАН,
г. Москва, Россия
kryukovkirill@yandex.ru
olkuznes@ipu.ru
suhoverv@ipu.ru

Предлагается метод алгоритмического установления компетентности научных сотрудников путем анализа содержания их публикаций. Метод основан на онтологии области научного знания. На основе этого метода осуществляется поиск экспертов, рецензентов, оппонентов, а также выбор команды для выполнения наукоемкого проекта.
Ключевые слова: компетентность, релевантность, онтология, терминология, рецензирование

Введение

Проблема определения компетентности стара как мир. Она, в частности, возникает всякий раз, когда требуется либо определить, пригоден ли данный человек для выполнения данной работы, либо выбрать из данной группы лиц человека, наиболее пригодного для выполнения данной работы. Для ее решения имеется множество приемов и процедур: заполнение анкет, тестирование, резюме, испытательный срок и т.д. Другой аспект этой проблемы – определение уровня владения определенными знаниями – всегда был существенной частью любого образовательного процесса. В последние десятилетия в рамках работ по управлению персоналом значительное внимание уделяется стандартизации и формализации процедур определения компетентности и попыткам разрабатывать информационные технологии, реализующие эти процедуры.

Очевидно, что для разных предметных областей методы определения компетентности имеют свою специфику. В настоящей работе рассматриваются задачи формализации некоторых процедур определения компетентности, типичных для научных организаций: алгоритмическое установление компетентности сотрудников на основе содержания их публикаций (компетентность, установленную таким методом, будем называть формальной); поиск экспертов, рецензентов, оппонентов, а также выбор команды для выполнения наукоемкого проекта на основе формальной компетентности.

1. Компетентность и релевантность

В данной работе мы будем исходить из стандартного определения компетенции как комбинации знаний и умений, позволяющей эффективно решать некоторый конкретный набор задач. В работе [Draganidis, 2006] компетенция описывается как следующая совокупность:

1) Наименование компетенции;
2) Категория, или область, к которой принадлежит компетенция;
3) Описание компетенции, объясняющее, что может делать обладатель компетенции;
4) Свидетельства компетенции, по наличию которых можно определить, обладает ли сотрудник данной компетенцией.

В дальнейшем мы будем различать компетенцию и компетентность. Компетенция – это понятие, в общем случае не связанное с конкретным лицом; компетентность – это отношение человек-компетенция, означающее, что человек владеет данной компетенцией. Соответственно, к понятию компетенции относятся пп. 1-3; п.4 должен подтверждать наличие отношения компетентности.

Существует множество работ по автоматическому определению компетентностей на основе документов. Упомянем в качестве примеров [Macdonald, 2006], [Ruger, 2009], [Ranwez, 2010]. Все они используют текстовый поиск на предмет

нахождения в документе свидетельств компетенции, которыми являются слова и словосочетания. В большинстве работ описанием компетенции служит запрос пользователя. Свидетельством обладания компетенцией является упоминание слов запроса в текстах документов, автором которых пользователь являлся. При этом могут использоваться методы расширения запросов, например, включение синонимов. Фактически этот подход представляет собой определение релевантности документа запросу.

Конкретизируем поставленную выше задачу определения научных компетенций в терминах приведенного выше определения компетенции.

Наименование компетенции совпадает с ее категорией – это некоторая область научного знания.

Под описанием компетенции мы будем понимать следующее. Формально описание компетенции представляется фрагментом онтологии, структурирующей научную область (скажем, теорию игр или науки об управлении) по отношению тема-подтема. Это формальное описание интерпретируется как владение тематикой, описанной выделенным фрагментом, достаточное для того, чтобы быть экспертом в данной области – рецензентом статей по этой тематике, оппонентом соответствующих диссертаций и т.д.

Источниками свидетельств компетентности будем считать публикации в рецензируемых журналах. Требование рецензируемости существенно, поскольку предлагаемый подход не ставит целью оценить качество публикаций (корректность, новизну, актуальность, стиль и т.д.). Поэтому предполагается, что качество уже положительно оценено рецензентами работы.

Возникает вопрос: что считать свидетельствами искомой компетенции в рассматриваемом документе? Выше отмечалось, что в большинстве предлагаемых методов таким свидетельством является достаточное число совпадений слов и словосочетаний в документе и в запросе. Однако для научных компетенций это не слишком удобно и к тому же не всегда дает нужный результат. Во-первых, в науке много компетенций, свидетельства которых не совпадают с названием. Например, в статье по тематике «Теория алгоритмов» слово «алгоритм» может ни разу не встретиться. Во-вторых, в названиях научных областей часто используются «неспециализированные» слова, которые можно встретить в статьях по любой тематике: тот же «алгоритм», «игра», «множество» и т.д. Научная тематика скорее характеризуется набором терминов, специфических для данной области. С другой стороны, вряд ли следует набор терминов всякий раз включать в текст запроса. Его тогда надо знать заранее – иначе разные пользователи, имея в виду одну и ту же тематику, будут включать в запрос разные термины. Поэтому целесообразно «объективировать» этот набор, т.е. включить его в описание предметной области.

Предлагаемый подход заключается в следующем. Предметная область научного знания представляется в виде онтологии с двумя типами вершин: вершины-темы и вершины-термины. Вершины-темы образуют основной каркас дерева и связаны между собой отношением тема-подтема, имеющим все таксономические свойства отношений типа класс-подкласс. Вершина-термин связана ровно с одной темой отношением тема-термин, но при этом предполагается, что все нижележащие вершины наследуют этот термин. Все термины являются висячими вершинами. На рисунок 1 приведен малый фрагмент онтологии теории игр, где термины изображены пунктирными прямоугольниками, а отношение тема-термин – пунктирными стрелками.

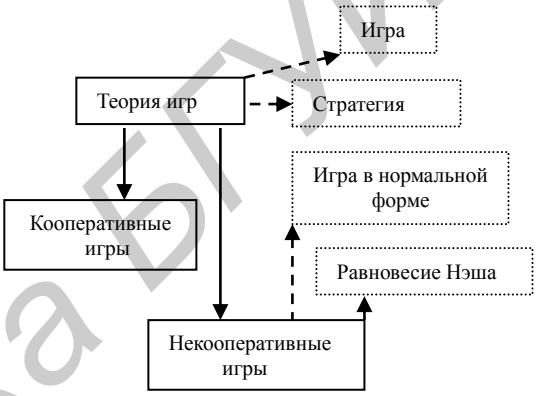

Рисунок 1 - Фрагмент онтологии теории игр

Релевантность документа теме характеризуется использованием основных терминов этой темы; компетентность сотрудника в теме – наличием публикаций, релевантных теме. И релевантность, и компетентность будем оценивать в непрерывной 5-балльной шкале. Интерпретация основных точек этой шкалы для релевантности – следующая.

0 – отсутствие релевантности.

1 – минимальная релевантность (в тексте есть хотя бы одно свидетельство темы);

2 – в тексте кратко упоминается тема (есть некоторое количество свидетельств темы, позволяющее оценить их расположение и разнообразие).

3 – в тексте часто упоминается тема, часть текста посвящена теме;

4 – в тексте достаточно подробно раскрыта тема, нераскрытыми остались лишь некоторые её части;

5 – тема рассмотрена максимально широко и разносторонне.

Интерпретация шкалы компетентности:

0 – отсутствие компетентности,

1 – минимальная компетентность,

2 – у автора есть некоторое представление о теме,

3 – автор работал с темой,

4 – автор хорошо знает тему и обладает достаточно широким кругозором в ней,

5- максимальная компетентность.

2. Формальное описание модели

Опишем предлагаемую модель формально.

Онтология предметной области строится так, как описано выше. Отметим нестандартность дерева онтологии: в нем висячие вершины (вершины-термины) имеются на всех уровнях иерархии. При этом отношение тема-термин не является таксономическим (типа часть-целое или класс-подкласс). Более естественно термины рассматривать не как экземпляры, а как атрибуты соответствующих тем.

Пусть C – множество всех тем онтологии; $c = |C|$; T - множество всех терминов онтологии.

Релевантность документа d темам онтологии определяется следующим образом.

Производится автоматическое аннотирование документа, т.е. поиск в нем всех терминов из T. В результате получается *аннотация* документа $V^d = \{v_1{}^d,\ldots, v_{|T|}{}^d\}$ –, где $v_j{}^d$ – число вхождений j-го термина в документе d.

Для определения уровня релевантности теме используются три параметра, которые вычисляются на основе найденных в документе терминов:

1) общее количество упоминаний основных терминов темы в документе;

2) объем темы - число фрагментов документа (например, абзацев или частей страницы), в которых встречались основные термины темы;

3) разнообразие основных терминов темы (число различных терминов) в документе.

Исходя из этих соображений, релевантность R_i^d документа d i-й теме вычисляется по следующей формуле

$$R_i^d = \alpha \cdot \frac{\sum_{n=1}^{|J_i|} v_n^d}{P_U} + \beta \cdot \frac{O_i^d}{P_O} + \gamma \cdot \frac{|J_i^d|}{|J_i|} \quad (1)$$

где α, β, γ - настроечные коэффициенты, $\alpha + \beta + \gamma = 5$; J_i – множество всех терминов, относящихся к теме i; $O_i{}^d$ – объем темы в документе; $J_i{}^d$ – множество всех различных терминов темы i, которые встретились в документе; P_U, P_O – нормирующие (так, чтобы соответствующие дроби не превосходили 1) величины для упоминаний и для объема, соответственно. Например, можно положить P_O равным O^d - объему всего документа. Объем темы измеряется с помощью понятия окна. Окно – это либо абзац, либо фиксированное число строк. Тогда $O_i{}^d$ – число окон, в которых встречается хотя бы один термин темы; O^d – общее число окон.

3. Профилем $PD(d)$ документа d называется его представление в виде вектора $PD(d) = \left(R_1^d,..R_c^d\right)$.

Формула релевантности для группы G документов обобщает формулу (1) и имеет следующий вид:

$$R_i(G) = \alpha \cdot \frac{\sum_{l=1}^{|G|}\sum_{n=1}^{|J_i|} v_n^{d_l}}{P_U} + \beta \cdot \frac{\sum_{l=1}^{|G|} O_i^l}{P_O} + \gamma \cdot \frac{|J_i^G|}{|J_i|} \quad (2)$$

где l – номер отдельного документа из коллекции документов; $J_i{}^G$ – множество всех различных терминов темы i, которые встретились в G. Нетрудно видеть, что, если G состоит из одного документа, то формула (2) переходит в формулу (1).

Будем считать, что уровень формальной компетентности автора в теме равен уровню релевантности всех его работ этой теме. Тогда профиль $PA(a)$ компетентностей автора a – это вектор релевантностей статей автора всем темам предметной области. Он имеет вид:

$$PA(a) = \left(R_1^{Da},..R_i^{Da}...\right),$$

где Da – все документы автора.

Следует иметь в виду, что предлагаемая оценка компетентности – это оценка снизу, т.е. «формальная компетентность», так как автор может иметь публикации не по всем темам, в которых он реально компетентен.

3. Приложения

Используя профили документов и сотрудников, можно решать задачи выбора экспертов (рецензентов, оппонентов) для оценки документа, консультантов по той или иной теме, а также подбора команды для работы над наукоемким проектом. Например, задача выбора рецензента решается в два этапа: сначала профили документа и кандидатов в рецензенты «очищаются» - в них обнуляются релевантности, которые ниже выбранного порога; затем профили кандидатов ранжируются по степени семантической близости к профилю документа с помощью одной из известных метрик (см., например, обзор [Крюков]).

В настоящее время проводится работа по созданию автоматизированной системы подбора рецензентов для журналов в области теории управления. Разработана онтология наук об управлении и словарь терминов с привязкой к терминам онтологии. Науки об управлении характеризуются большим разнообразием прикладных теорий, математических методов и областей приложения. Поэтому верхний уровень онтологии выглядит следующим образом (рисунок 2):

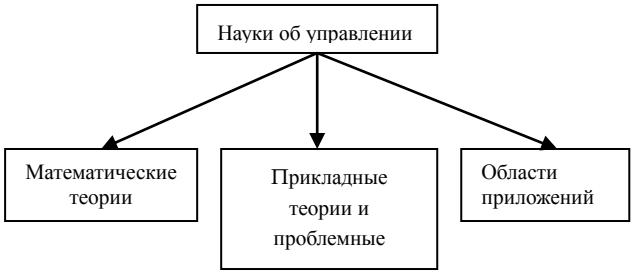

Рисунок 2 – Верхний уровень онтологии

На следующем уровне иерархии под «Математическими теориями» располагаются вершины «Алгебра», «Дифференциальные уравнения», «Математическая логика», «Теория вероятностей» и др., под «Прикладными теориями» – вершины «Теория автоматического управления», «Автоматизация проектирования», «Анализ данных», «Искусственный интеллект», «Передача и обработка сигналов», «Теория организационных систем», под «Областями приложений» - вершины «Вычислительные и коммуникационные системы и сети», «Социально-экономические системы», «Бизнес и финансы», «Технологические процессы», «Робототехника», и др.

Словарь терминов получен путем анализа статей журнала «Автоматика и телемеханика» за последние два года; в дальнейшем он будет пополняться. При формировании словаря важно было выбирать термины, характерные для данной темы; поэтому такие часто встречающиеся, но «безликие» термины, как «функция», «система», «уравнение», в словарь не вошли. Ясно, что такой отбор терминов можно было провести только экспертным путем.

Заключение

В работе предложен метод автоматического определения компетентности научных сотрудников путем анализа содержания их публикаций. Метод основан на специальной онтологии области научного знания, ядром которой является иерархия тем научной области, связанных таксономическим отношением тема-подтема. Каждой теме соответствует набор терминов, ее характеризующих. Релевантность документа (публикации) теме вычисляется по пятибалльной шкале, исходя из трех параметров: общее количество упоминаний основных терминов темы в документе; число фрагментов документа (например, абзацев или частей страницы), в которых встречались основные термины темы; разнообразие основных терминов темы (число различных терминов) в документе. Профиль документа определяется как вектор его релевантностей всем темам онтологии. Профиль компетентности сотрудника определяется как профиль группы его публикаций. Показано, как, используя профили документов и сотрудников, осуществлять выбор рецензентов статьи, консультантов по теме и подбор команды для работы над наукоемким проектом.

Работа выполнена при поддержке РФФИ (грант № 11-01-00771).

Библиографический список

[Draganidis, 2006] Draganidis F., Mentzas G. Competency based management: a review of systems and approaches/ Information Management & Computer Security, 2006, Vol. 14, No. 1, pp. 51-64.

[Macdonald, 2006] Macdonald C., Ounis, I. Voting for candidates: adapting data fusion techniques for an expert search task/ CIKM 2006: Proc. of the 15th ACM International Conference on Information and Knowledge Management. ACM, New York (2006), pp. 387–396.

[Ranwez, 2010] Ranwez Sylvie et al. User Centered and Ontology Based Information Retrieval System for Life Science/ Available from Nature Precedings, 2010. http://precedings.nature.com/documents/5408/version/1/files/npre20105408-1.pdf (28/03/2011).

[Ruger, 2009] Ruger S., Zhu J., Song D. Integrating multiple windows and document features for expert finding/ Journal of the American Society for Information Science and Technology, 2009, 60 (4), pp. 694-715.

[Крюков, 2010] Крюков К.В, Панкова Л.А., Пронина В.А., Суховеров В.С., Шипилина Л.Б. Меры семантической близости в онтологиях// Проблемы управления, 2010, т. 5, с. 2-14.

A CONCEPT OF FORMAL COMPETENCE FOR RESEARCH WORKERS

Kryukov K.V., Kuznetsov O.P., Suhoverov V.S.

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

kryukovkirill@yandex.ru

olkuznes@ipu.ru

suhoverv@ipu.ru

The paper describes the method of competence calculation for research workers by content analysis of their publications. The method uses the special ontology of science domain. The application of this method for a selection of reviewers, opponents and a team for high-tech project is shown.