УДК 004.82

ОНТОЛОГИИ И ПЕРСОНИФИКАЦИЯ ПРОФИЛЯ ПОЛЬЗОВАТЕЛЯ В РЕКОМЕНДУЮЩИХ СИСТЕМАХ ТРЕТЬЕГО ПОКОЛЕНИЯ

В.И. Городецкий[1], О.Н. Тушканова[2]

Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, Россия
[1]gor@iias.spb.su, [2]tushkanova@iias.spb.su

Аннотация

Рекомендующие системы третьего поколения сейчас находятся в самом начале своего развития. К рекомендующим системам третьего поколения относят системы, которые вырабатывают решения на основе семантических моделей интересов и предпочтений пользователя, принимают во внимание мотивацию и причины, которые побуждают конкретного пользователя предпочитать то или иное решение, а также учитывают семантику контекста, сопутствующего принятию решений. Иначе говоря, основным новшеством таких систем является их ориентация на семантические модели представления и использования знаний, в частности, знаний о персональном профиле пользователя. В настоящее время наиболее естественным и наиболее разработанным способом формализации семантических категорий, которые обычно используются человеком в процессах выработки решений, является онтология. По этой причине в современных и будущих системах, которые называют рекомендующими системами третьего поколения, онтология рассматривается в качестве общей структуры для представления разнообразных и разнотипных знаний. Примерами таких знаний являются, например, знания о персональном профиле пользователя, о контексте принятия решений, об эмоциональном состоянии пользователя при принятии решения и т.п. Данная работа посвящена обзору публикаций, которые рассматривают подходы и модели персонального профиля пользователя в контексте онтологий предметной области рекомендаций, а также варианты их использования в рекомендующих системах. Эти публикации фактически подготовили современный взгляд на онтологию как единую модель концептуализации и представления компонент знаний, используемых в рекомендующих системах.

Ключевые слова: *рекомендующие системы, онтология, профиль пользователя, персонификация, контекстно-зависимые рекомендации, фольксономия, сетевой контекст, автоматизация разработки онтологий.*

Введение

Концепция рекомендующих систем как систем прогнозирования решений человека при выборе им товара/услуги появилась в середине 1990-х годов. К настоящему времени эта концепция прошла стандартный путь развития, характерный для систем, основанных на знаниях. Она прошла путь от статистических и эвристических моделей обучения и принятия решений к моделям, в которых используются содержательно интерпретируемые знания, учитывается контекст принятия решений, а также персональные интересы и предпочтения пользователя, чье решение рекомендующая система пытается прогнозировать.

В настоящее время предметом перспективных исследований в этой области являются рекомендующие системы третьего поколения [1], которые акцентируют внимание на семантических моделях представления и использования всех компонент знаний, которые вовлекаются в процесс выработки рекомендаций. С функциональной точки зрения, рекомендующие системы третьего поколения должны вырабатывать решения на основе семантических моделей интересов и предпочтений пользователя, принимать во внимание мотивацию и причины, которые побуждают конкретного пользователя делать тот или иной выбор. Они должны так-

же учитывать семантику контекста, сопутствующего принятию решений пользователем. Системы третьего поколения должны быть способны давать мотивированные объяснения предлагаемым решениям. Они должны учитывать экономические, психологические и другие факторы, связанные с принятием решений [1].

Рекомендующие системы третьего поколения сейчас находятся в самом начале своего развития. Потребовалось около 15 лет для того, чтобы сформировать современную концепцию рекомендующих систем. В ней центральной идеей является использование семантических категорий естественного языка для концептуализации, представления и использования знаний, вовлекаемых в процессы выработки решений. В настоящее время наиболее естественной и наиболее разработанной моделью формализации семантических категорий, которые используются человеком, является модель онтологии. Поэтому в рекомендующих системах третьего поколения онтология рассматривается в качестве общей модели концептуализации и структуры для представления всех компонент знаний, например, знаний о предметной области рекомендаций, знаний о персональном профиле пользователя, о контексте принятия решений, об эмоциональном состоянии пользователя при принятии решения и т.п.

Представление профиля пользователя в виде структур на множестве понятий предметной онтологии становится особенно продуктивным и экономным, если эти понятия являются *причинами*, которые определяют его отношение к выбору того или иного товара/услуги. Такие структуры знаний однозначно и естественным образом задают механизмы принятия решений пользователя, представленные в декларативной форме. Их принято называть активными структурами знаний, или, другими словами, знаниями, которые *способны генерировать действия* (англ. *actionable knowledge*[1]) [2].

Данная работа посвящена обзору публикаций, в которых предлагаются онтологические модели интересов пользователя. Эти публикации фактически подготовили современный взгляд на онтологию как единую модель концептуализации и представления компонент знаний, используемых в рекомендующих системах. В разделе 1 даются краткие сведения о рекомендующих системах в контексте их развития, а также сведения о «классических» методах прогнозирования решений пользователя. Раздел 2 поясняет понятие профиля пользователя и его связь с онтологией предметной области рекомендаций. Раздел 3 описывает те модели представления профиля пользователя, которые непосредственно предшествовали онтологическим моделям, однако во многом способствовали принятию этой концепции для перспективных рекомендующих систем. В разделе 4 рассматриваются модели профиля пользователя, которые уже полностью базируются на использовании онтологий.

Следует заметить, что разработка онтологической модели предметной области рекомендаций и, в частности, онтологической модели профиля пользователя требует решения ряда новых специфических проблем, с которыми разработчики онтологий ранее не сталкивались. Это связано с тем, что при построении модели предметной области рекомендаций и профиля пользователя приходится использовать множество разнородных источников информации и данных, в которых, так или иначе, отражаются «следы» присутствия или деятельности пользователя. Эти источники обычно распределены по сети Интернет и могут содержать весьма специфичную информацию. К такой информации относятся, например, тэги (англ. *tags*). Тэги – это семантически осмысленные метки. Ими сам пользователь или другие пользователи обычно метят продукты/услуги, которые могут быть или уже были *объектами выбора* (англ. *items*) тех или иных пользователей. Информация о тэгах обычно является достаточно информативной и семантически богатой. Она особенно информативна для задачи персонификации интересов и предпочтений пользователя. Тэги как элементы естественного языка обычно с

[1] “Propositions that are actionable are those that actors can use to implement effectively their intentions. Actionable knowledge requires propositions that make explicit the causal processes required to produce action. Causality is the key in implementation” [3].

трудом отображаются на понятия, которые имеются в стандартных словарях понятий. Поэтому категоризация тэгов – это весьма нетривиальная задача, которую, однако, нужно решать. Обычно тэги связываются с одним или несколькими понятиями и категориями в специальных словарях, которые принято называть *фольксономиями*. Проблема построения профиля пользователя на основе тэгов рассматривается в разделе 5, в котором также даются краткие сведения о существующих подходах к построению онтологий на основе тэгов. Еще одна специфическая проблема возникает при построении онтологического профиля пользователя в случае рекомендаций, формируемых в контексте социальных сетей. Особенности построении онтологического профиля пользователя и возможные способы учета «сетевого контекста» демонстрируются на примере в разделе 6. Во многих случаях онтологию предметной области рекомендаций приходится строить с использованием множества источников данных. Краткие сведения о предложенных подходах к построению онтологий и профиля пользователя для таких случаев приведены в разделе 7.

В заключении по работе приводится краткая сводка преимуществ онтологических моделей профиля пользователя, а также перечисляются некоторые тенденции в этой области. Кроме того, дополнительно дается краткий перечень проблем, не затронутых в данном обзоре, которые, однако, важны и имеют свою специфику применительно к построению профиля пользователя в рекомендующих системах и требуют специального рассмотрения.

1 Краткие сведения о рекомендующих системах

Рекомендующими системами[2] называют класс систем принятия решений, которые используют знания об интересах и предпочтениях человека для оценки/прогнозирования его реакции на рекомендацию купить некоторый товар или воспользоваться некоторой услугой [4]. Потенциальная реакция человека оценивается системой с помощью некоторой метрики, которая измеряется обычно в порядковой шкале, например, в целочисленной шкале. Эта метрика характеризует меру уверенности системы в том, что человек отреагирует на ее рекомендацию положительно. Ее принято называть *рейтингом* товара/услуги.

Исследования в области рекомендующих систем начались в середине 1990-х годов. Появление и развитие систем такого типа в значительной степени было обусловлено задачами маркетинга. В настоящее время интерес к практическому использованию рекомендующих систем постоянно возрастает, и главной, хотя и не единственной причиной этого интереса является повышение относительной доли Интернет-торговли в общем объеме продаж товаров и услуг. И это понятно, поскольку потенциальный покупатель товара или потребитель услуги в Интернет-магазине не видит сам товар. При отсутствии рекомендаций, соответствующих его интересам и предпочтениям, он вынужден был бы искать нужное в толстых каталогах или базах данных о товарах или услугах конкретного магазина, отслеживать появление новых товаров и т.д. Очевидно, что ни один нормальный покупатель товара или потребитель услуги не стал бы на это тратить так много времени, а пошел бы в обычный магазин, где все товары выставлены, или в офис, в котором всю информацию о потребной услуге можно быстро получить.

К настоящему времени рекомендующие системы получили уже значительное теоретическое развитие, и нашли широкое практическое применение. Рекомендующие системы первого поколения (1995-2005+), которые и сейчас имеют широкое хождение, использовали модель, основу которой составляли три матрицы [1]. Первая из них, матрица $\boldsymbol{X} = \{x_{ij}\}_{i-1,\,j=1}^{M,\,P}$,

[2] В научной литературе на русском языке равноправно используются два термина: *рекомендующие* (англ. *recommending*) системы и *рекомендательные* (англ. *recommender*) системы. Читатель вправе сделать контекстную замену первого термина, который и используется в работе, на второй, если ему так кажется лучше.

задает по строкам множество имен пользователей $C=(c_1,...,c_M)$, а по столбцам— множество имен свойств (атрибутов) пользователей $X=(x_1,x_2,...,x_P)$, так что элемент x_{ij} матрицы $\boldsymbol{X}$ задает значение атрибута x_j для пользователя c_i. Вторая матрица, $\boldsymbol{Y}=\{y_{ij}\}|_{i=1,j=1}^{N,Q}$, задает по строкам множество имен s_i товаров/услуг $S=(s_1,...,s_N)$, а по столбцам – множество $Y=(y_1,...,y_Q)$ атрибутов (свойств) y_j этих товаров/услуг. Наконец, третья матрица $\boldsymbol{R}=\{r_{i,j}\}|_{i=1,j=1}^{M,N}$, которая называется *матрицей рейтингов*, имеет в качестве строк имена пользователей $C=(c_1,...,c_M)$, в качестве столбцов – имена товаров/услуг $S=(s_1,...,s_N)$. Ее элементом $r_{i,j}$ является значение рейтинга товара/услуги s_j, присвоенное ему пользователем c_i. Обычно матрица рейтингов $\boldsymbol{R}$ бывает слабо заполненной, поскольку общее число товаров/услуг всегда значительно больше, чем то их число, которым пользователь реально мог воспользоваться и оценить с помощью рейтинга. Задача рекомендующей системы состоит в том, чтобы спрогнозировать значения рейтинга, который будет присвоен пользователем новому для него товару/услуге в матрице $\boldsymbol{R}$. Обратим внимание на то, что интересы пользователя здесь представлены в матричной форме, а в модели знаний рекомендующей системы онтология не используется.

В рекомендующих системах первого поколения используются разные методы принятия решений, однако базовыми являются три из них [1]:

- Методы, основанные на фильтрации контента (множества свойств) $Y=(y_1,...,y_Q)$, описывающего продукты/услуги, которые потенциально могут быть рекомендованы пользователю. В этих методах прогноз пользовательского рейтинга нового товара/услуги формируется по его сходству с другими товарами/услугами, которые пользователь ранее оценил с помощью рейтинга («скажи мне, какие рейтинги данный пользователь присваивал раньше похожим товарам/услугам, и я скажу тебе значение рейтинга нового товара для него» [5]). Для алгоритмической реализации этого метода должна быть задана форма метрики, по которой вычисляется сходство товаров/услуг. Обычно для этих целей используются различные метрики сходства [6].
- Методы коллаборативной фильтрации, в которых полагается, что текущие предпочтения пользователя будут сходны с предпочтениями других пользователей с похожими интересами по отношению к предлагаемому товару или услуге. В этом случае решающую роль играет сходство пользователей по своим предпочтениям («скажи мне, какой рейтинг присвоили этому товару другие пользователи, интересы которых сходны с интересами данного пользователя, и я скажу тебе, какой рейтинг он присвоит этому товару/услуге» [5]).
- Гибридные методы, которые комбинируют оба названных выше подхода и, возможно, используют еще какие-то новшества.

Важно отметить, что по своей сути все эти методы являются чисто статистическими и весьма слабо учитывают персональные предпочтения конкретного пользователя. Пользователь в них характеризуется выборкой примеров товаров/услуг, которым он присвоил значения рейтинга в той или иной шкале. По этой причине предсказательная сила рекомендующих систем первого поколения оказалась достаточно слабой.

Отличительной чертой рекомендующих систем второго поколения (2003-2014+) является, в первую очередь, учет контекста, который сопутствует рекомендациям. В качестве главных измерений (атрибутов) контекста рассматриваются обычно время, место и социальный контекст. Например, если пользователю рекомендуется фильм, то следует учитывать желаемое время просмотра (рабочий день или выходной день и т.п.), хочет ли пользователь смот-

реть фильм дома или в кинотеатре и, наконец, в каком коллективе он планирует просмотр (с родителями, с друзьями и т.п.). Детальный анализ сущности контекста, моделей его формального представления и учета в процессах выработки рекомендаций можно найти в [7–9], а также в [4].

Рекомендующие системы второго поколения характеризуются и другими возможностями, которыми их предшественники не обладали. К ним относится, например, использование многокритериальных рейтингов и рейтингов, формируемых сразу для групп пользователей. Для оценки товаров/услуг в них можно использовать не только рейтинг, но и простое упорядочение рекомендуемых вариантов выбора. Они способны вырабатывать рекомендации в контексте социальных сетей, использовать тэги в качестве источника знаний и профиле пользователя и т.п. Обобщая сказанное, можно утверждать, что в системах второго поколения появился акцент на использование *знаний об интересах* пользователя и о *контексте конкретной задачи* выработки рекомендаций. Заметим, что для формализации знаний о профиле пользователя в рекомендующих системах второго поколения начали использовать их представление в терминах онтологий[3].

В настоящее время наиболее передовые *практические* разработки в области рекомендующих систем соответствуют второму их поколению, и активные научные исследования в интересах развития таких систем продолжаются. Однако для исследователей основной научный интерес представляют уже рекомендующие системы третьего поколения [1]. Такие системы, прежде всего, ориентируются на семантические аспекты моделей интересов и профиля пользователя. В них большое значение придается мотивационным аспектам того или иного выбора пользователя и причинам, которые определяют его выбор. Полагается, что рекомендующие системы третьего поколения должны быть способными давать мотивированные объяснения решениям, которые они предлагают. Считается, что рекомендующие системы третьего поколения должны учитывать экономические, психологические и другие факторы, определяющие выбор пользователя [1]. Очевидно, что большинство перечисленных свойств носит семантический характер, акцентирует внимание разработчика на представлении модели интересов и предпочтений пользователя в терминах семантических категорий, обычно используемых человеком в естественном языке.

Такая тенденция наметилось еще в период развития систем второго поколения, однако в системах третьего поколения эта идея становится основной. Поэтому онтологии стали постепенно рассматриваться в них в качестве семантической концепции и структуры для представления всех компонент знаний. В этой концепции онтология должна объединять в себе знания о предметной области рекомендаций и о персональном многомерном профиле пользователя, поддерживать процессы выработки кросс-доменных рекомендаций (например, имея информацию об интересах пользователя в области музыки, формировать рекомендации в области кинофильмов или художественной литературы). Контекстная зависимость профиля интересов пользователя в онтологической концепции рекомендующих систем также должна быть представлена в онтологической модели знаний.

Динамика формирования этой концепции, предложенные подходы к представлению в ней персонального профиля пользователя и ее современное состояние рассматриваются в последующем материале работы.

2 Профиль пользователя

Построение персонального профиля пользователя в течение последнего десятилетия является одной из самых «горячих» тем исследований и разработок в области рекомендующих

[3] Естественно, что провести *точные* границы между рекомендующими системами разных поколений невозможно.

систем, и она остается такой для рекомендующих систем третьего поколения. Базовая идея онтологической модели формализации профиля пользователя состоит в том, что понятие *интерес пользователя* рассматривается как экземпляр некоторой категории в иерархии категорий (понятий) онтологии, а *профиль пользователя* – как некоторая структура, заданная на множестве интересов-категорий онтологии, возможно, структурированных иерархически. Поэтому *построение онтологии* и *выделение в ней профиля пользователя* в настоящее время являются двумя тесно связанными задачами разработки рекомендующих систем. Обычно эта связь обусловлена тем, что и для построения онтологии предметной области рекомендаций, и для построения профиля пользователя используются одни и те же источники данных.

Если онтология предметной области построена, то поиск персонального профиля пользователя сводится к поиску множества его интересов, выраженных в терминах понятий этой онтологии или их примеров. Это может выполняться различными алгоритмическими средствами. Выбор зависит и от объема и состава исходной информации об интересах пользователя, и от метода, который далее предполагается использовать для принятия решений, и от других факторов.

Методы построения профиля пользователя варьируются от простого опроса самого пользователя, до полностью автоматического (машинного) обучения профиля путем обработки всей доступной информации, полученной из разных источников, в которой, так или иначе, проявляются «следы» поведения пользователя, отражающие его интересы, предпочтения, намерения, контекст принятия решения и прочее.

Далее приводится обзор результатов в области построения персонифицированного профиля пользователя в контексте онтологии предметной области рекомендаций для рекомендующих систем третьего поколения.

3 Онтологии и персонификация профиля пользователя: начальные модели

Идея построения персонального профиля пользователя с использованием категорий или понятий естественного языка была первоначально предложена для рекомендующих систем в области поиска документов в Интернет [10–13]. Эти методы принято называть методами на основе понятий (англ. *concept-based methods)*. По существу, эти работы еще не использовали в полной мере онтологические модели профиля пользователя, однако они в последующем стимулировали их развитие. Общая идея этой группы методов состоит в том, чтобы найти в текстах документов, выдаваемых поисковой машиной в ответ на запросы пользователя, слова, которые могут рассматриваться как понятия. При этом принимается гипотеза, что эти слова–понятия отражают, с одной стороны, семантику запроса пользователя, а с другой – персональные интересы пользователя. Хотя эти подходы и напоминают традиционный подход на основе ключевых слов, однако, они не являются таковыми в полной мере. Первый шаг в этих подходах состоит в поиске часто встречающихся терминов с использованием традиционной меры, применяемой при поиске ключевых слов с помощью методов обнаружения часто встречающихся паттернов. Для того чтобы быть более уверенными в том, что термины, выделенные таким способом, не являются просто ключевыми словами, далее тем или иным способом выполняется их сравнение с понятиями некоторой базы данных понятий, выбранной разработчиком. Например, работа [12] формирует разделы интересов пользователя по множеству веб-документов его истории поиска и использует базу данных понятий *Open Directory Project* (*ODP*) [14]. Далее профиль пользователя представляется в виде множества найденных категорий с *весами*, пропорциональными частоте их появления в документах. Важно отметить, что далее делается шаг обобщения: каждая категория сопровождается списком ключевых слов, встречающихся в документах, которые могут рассматриваться в каче-

стве *примеров* этого понятия. Аналогичный подход используется в работе [10]. Преобразование найденных терминов в онтологию выполняется с помощью специального механизма классификации, построенного авторами. В работе [13] дополнительно к документам, которые ранее просматривались пользователем, используются также письма его электронной почты. В них находятся часто встречающиеся термины, и при этом так же, как и в ранее названных работах, полагается, что они отражают интересы пользователя в содержании веб-документов, которые он запрашивал. В качестве базы данных понятий в этой работе, как и в работе [12], используется структура категорий *ODP* [14].

Авторы работы [11] указывают в качестве общего недостатка работ [12–13] *статический* характер профиля пользователя, который строится с использованием практически не обновляемой базы данных *ODP* [14]. Новым в работе [11] является *персонификация каждого запроса* пользователя. И это очень важно для поиска веб–документов, так же, как и в других задачах поиска *персонифицированных* ответов на запросы. Для достижения этой цели авторы анализируют преамбулу (служебную информацию) каждого документа, возвращаемого на запрос пользователя поисковой машиной, называемую *Web-snippet*. Информация преамбулы включает в себя название документа, аннотацию его содержания, а также URL–адрес веб-документа, возвращаемого поисковой машиной в ответ на запрос. Далее авторы полагают, что те ключевые слова, которые часто встречаются в служебной информации возвращаемых документов, являются искомыми понятиями, которые отражают интересы пользователя. Для отбора таких понятий авторы, как и большинство авторов других работ, используют пороговое значение меры, которая обычно используется в алгоритмах фильтрации (отбора) часто встречающихся паттернов. Эта мера вычисляется как произведение числа слов в понятии и частоты встречаемости понятия (число преамбул, в которых встречается это понятие, деленное на общее число возвращенных документов или проанализированных преамбул). Дополнительным новым элементом работы [11] является использование как *позитивных*, так и *негативных* интересов. По каждому запросу авторы сначала извлекают множество понятий на основе служебной информации, затем выполняется кластеризация найденных понятий по некоторой мере сходства и построение на этой основе персонифицированного профиля пользователя. Авторы рассматривают несколько таких стратегий, предложенных в различных их работах.

Достоинство проанализированной группы работ состоит в том, что они отказываются от простого использования ключевых слов, объединяя их в понятия. Они привлекают дополнительную информацию (например, базу данных *ODP* или аналогичную по содержанию другую базу данных понятий) для поиска понятий, представляющих персонифицированный профиль пользователя в форме множества найденных понятий с приписанными им весами. Другое достоинство этих работ – это использование сжатой информации о веб-документе, содержащейся в его преамбуле.

Однако эти работы делают лишь первый шаг от традиционных стратегий, которые используют в качестве атрибутов ключевые слова, к стратегиям на основе онтологий. Они используют только частотные характеристики для взвешивания интересов в профиле пользователя, но не анализируют зависимости между ними и не устанавливают отношений на множестве извлеченных терминов, которые рассматриваются далее как понятия. Бинарная постановка задачи, в которой не используется понятие рейтинга документа с точки зрения пользователя, обычно обеспечивающего обратную связь от него, не позволяет корректировать значимость различных интересов в профиле пользователя, и это сильно ограничивает возможности описанной группы методов.

Другие работы периода 2004–2008 гг., аналогичные по используемой идее[4], предлагающие «почти» онтологические модели персонифицированного профиля пользователя на основе понятий, как правило, посвящены разработке конкретных приложений, а потому сильно ориентированы на их особенности. Примерами таких работ являются [15–17] и ряд других.

В [17] используется классификация веб-страниц. Основное внимание в ней уделяется персонификации профиля пользователя, представленного терминами *выбранной заранее* онтологии, в данном случае, в терминах *иерархии ODP* [14]. В качестве исходной информации для моделирования профиля пользователя используются веб-страницы (документы), классифицированные вручную по типу интереса для пользователя таким образом, чтобы каждый построенный класс отражал некоторый его конкретный интерес. Множество документов выборки, поставленной в соответствие одному классу, авторы называют «супердокументом» класса, который далее используется в качестве обучающей выборки этого класса. Каждому классу ставится в соответствие то, что авторы называют понятием (по существу – это некоторый интерес пользователя). Процесс обучения состоит в том, что каждому понятию-классу далее ставится в соответствие вектор терминов, сформированных с помощью ключевых слов и словаря понятий *ODP* [14] с весами, которые вычисляются по «супердокументу» класса. В итоге каждое понятие-класс представляется вектором терминов из словаря *ODP* с весами.

Классификация веб-страниц проводится на основе вычисления меры сходства вектора весов терминов, созданного для веб-страницы, с аналогичным вектором каждого понятия-класса, причем в качестве меры сходства выбрано значение косинуса угла между этими векторами. Для поиска лучших совпадений используется также метод *К–ближайших соседей*.

Работа [16] предлагает онтологическую модель профиля пользователя в онлайн-системе рекомендаций научно-исследовательских работ. Эти работы классифицируются с использованием онтологических классов и механизма коллаборативной фильтрации (англ. *collaborative filtering*). Профиль интересов пользователя представляется в понятиях онтологии научной области исследовательских работ, разработанной в рамках инициативы АКТ [18]. Связи, присутствующие в этой онтологии, используются для того, чтобы выводить другие понятия-темы, которые могут представлять интерес для пользователя, в том числе и те, которые еще не были им просмотрены.

В [15] предлагается модель профиля пользователя, охватывающая несколько предметных областей, представленных единой моделью онтологии. Эта работа имеет целью преодолеть ограниченность работ, в которых профиль пользователя представлен ключевыми словами, не согласованными между собой семантически, т.к. последнее может приводить к противоречивым моделям. Кроме того, авторы встраивают в модель онтологии контекст, сопровождающий процесс выработки рекомендаций. В работе предлагается рекомендующая система *CORеS* (*Context-aware Ontology-Based Recommender System*), использующая онтологию. Эта система способна работать с несколькими предметными областями и может привлекать контекст для получения более точных персонифицированных рекомендаций. «*Еще одно новшество CORеS– разбиение профиля пользователя на части в соответствии с различными областями рекомендаций, которое формируется в контексте решаемой задачи*», – отмечают авторы. К сожалению, в работе рассматриваются аспекты мульти-доменного профиля пользователя, но не кросс-доменного профиля. Последнее было бы значительно интереснее, поскольку является гораздо более востребованным на практике.

Далее рассматриваются работы такого же направления, которые ориентированы уже на полноценное использование онтологий предметной области рекомендаций и ее категорий для представления персональных пользовательских интересов. Эти работы активно исполь-

[4] В этот период появились также работы, которые использовали уже полномасштабные онтологические модели для представления профиля пользователя.

зуют также различные существующие общедоступные онтологии и базы знаний для того, чтобы *автоматизировать* процессов построения онтологических профилей пользователей.

4 Онтологии и персонификация профиля пользователя: современные модели

Модели, которые в полной мере используют возможности онтологии как структуры для представления знаний о предметной области рекомендаций для представления профиля пользователя в терминах ее понятий, активно развиваются в течение последних примерно 6 лет. Важно отметить, что ключевым фактором во всех этих исследованиях является состав и свойства исходной информации, которая доступна для построения онтологии.

Другой общей чертой этих работ является поиск путей *автоматизации* процессов, связанных с разработкой персонального профиля пользователя в контексте предметной онтологии. Для этих целей используются различные словари понятий, стандартные онтологии повторного использования, викификация[5] веб данных [19], развиваемые в настоящее время параллельно со всем хорошо известным вебом документов, и другие источники знаний, стандартные программные средства и программные платформы.

В работе [20] представлена частная технология автоматизированного построения формальной модели персонального профиля пользователя в терминах онтологии, характеризующей его интересы и намерения. Она ориентирована, главным образом, на пользователей, работающих в социальной сети *Twitter*, например, для последующей рекомендации друзей в этой сети или для рекомендации им товаров и услуг с помощью более целенаправленной рекламы.

В качестве входной информации в разработанной модели используются URL-адреса, извлеченные из постов пользователя в *Twitter* (из «твитов»). Верхний уровень онтологии, используемой авторами, который отвечает категоризации веб-сайтов, формируется вручную с использованием понятий, извлеченных из дополнительных источников. Примерами таких источников являются базы коллективных знаний *OpenDNS* [21], таксономии рекламных объявлений и онтологии *DBpedia* [22]. *OpenDNS* накапливает концептуализацию *тэгов* и реализуется как особый облачный сервис, который позволяет членам сообщества пополнять множество пар *тэг – категория* веб-сайта и относить веб-сайты к некоторым категориям, например, *Музыка, Кинофильм* и т.п. Категории, к которым может относиться веб-сайт, должны быть одобрены сообществом, прежде чем они попадут в эту базу данных. *DBpedia* извлекает категоризацию из дампа категорий Википедии и преобразует их представление в *RDF*-формат. *OpenDNS* и *DBpedia* покрывают целый ряд предметных областей, т.е. они позволяют строить мульти-доменные онтологии интересов пользователей.

Для построения модели профиля пользователя авторы представляют онтологию на языке *OWL* или *RDF/XML*. Для хранения твитов, содержащих URL, и сформированных профилей пользователя в формате *RDF* авторы используют *NoSQL*-базы данных *Cassandra* [23] и *Virtuoso* [24], соответственно. Заметим, что система управления базами данных *Cassandra* представляет собой распределённую *NoSQL*-систему [25], рассчитанную на создание масштабируемых и надёжных хранилищ огромных массивов данных, представленных в виде хэша. *Virtuoso Universal Server* является системой с открытым исходным кодом, которая сочетает в себе возможности реляционной базы данных, объектно-реляционной базы данных и вирту-

[5] Так принято называть технологию автоматизации построения онтологий, в которой используется дамп категорий Википедии в качестве словаря понятий с возможностью получения статей, представляющих семантику этого понятия. Технология предложена И. Виттеном (I. Witten) из Университета Вайкато, Новая Зеландия, в 2007 году.

альной базы данных. Она хранит данные в формате *RDF* и *XML*. Она может выполнять функции сервера веб-приложений и файлового сервера.

Следует отметить, что использование *NoSQL*-баз данных позволяет улучшить масштабируемость приложений и обеспечивает возможность работы с большими данными, в частности, поступающими в реальном времени, что сегодня является обязательным требованием к системам, работающим с социальными сетями. Заметим, что использование *NoSQL*-баз данных отражает современную тенденцию рассматривать рекомендующие системы как системы, функционирующие на основе анализа больших данных.

Рассмотрим процедуру формирования профиля пользователя, предложенную в работе [20]. На первом этапе URL-адреса извлекаются из твитов и помещаются в базу данных *Cassandra*. Авторы полагают, что *все* URL, размещенные пользователями в твитах, представляют его интересы. С этим нельзя согласиться безоговорочно, поскольку пользователи могут делиться не только тем, что их интересует, но и чем-то, что вызывает у них негативные эмоции, или же размещать в твите какую-либо информацию по просьбе своих друзей или в рекламных целях. При этом понятие «негативного» интереса или интересов с присвоенными значениями рейтингов авторами в работе вообще не рассматривается. Далее, анализ сопутствующих текстов, а также текста, содержащегося в твите, не выполняется. Вообще говоря, это может приводить к потере существенной информации.

Твиты, помещенные в базу данных *Cassandra*, постоянно анализируются в пакетном режиме и обрабатываются с целью извлечения интересов и намерений пользователя (см. далее по тексту, как это выполняется), которые затем добавляются в его онтологический профиль, хранимый в базе данных *Virtuoso* в виде *RDF*-триплетов. Для поиска дополнительных интересов в существующей онтологии профиля пользователя используется механизм рассуждений *Pellet* [26].

Алгоритм извлечения интересов и намерений пользователя состоит в следующем. Сначала все URL-адреса пользователя разделяются на *интересные* и *неинтересные*. К «неинтересным» авторы относят URL-адреса поисковых машин (*Google, Yahoo* и т.п.), сервисы сокращения ссылок (*Goo, Su* и т.п.), URL социальных сетей (*Facebook, Twitter* и т.п.). Далее анализируются только «интересные» URL-адреса. Отметим, что такую фильтрацию (очистку, редукцию исходных данных) нельзя отнести к бесспорным преимуществам алгоритма, поскольку таким способом можно удалить большой объем полезной информации о пользователе.

Далее, для каждого URL-адреса из баз коллективных знаний *OpenDNS* и *DBpedia* извлекаются категории и понятия, которые добавляются к онтологии профиля пользователя, вместе с отношениями *hasIntent* или *hasInterest*. Дополнительно извлекаются категории, связанные отношением *sameAs* с категориями URL, для которых в онтологии пользователя создается отношение *categorySameAs*.

Фрагмент онтологии, разработанной в [20], в части, касающейся представления профиля пользователя, изображен на рисунке 1. В нем понятие *Персона* – класс, который идентифицирует пользователя; URL – класс, включающий в качестве примеров URL-адреса, размещенные в твитах; *Интерес* – класс, который представляет понятия, связанные с понятием URL; *Намерение* – класс, который представляет понятия, связанные с понятием URL, выражающие покупательские намерения пользователя; *Неизвестно* – класс, содержащий URL-адреса, которые не существуют в базах данных коллективных знаний (*OpenDNS, DBpedia*); *НеизвестнаяКатегория* – класс категорий, которые не существуют в БД коллективных знаний (*OpenDNS, DBpedia*).

Авторы определяют несколько подклассов классов *Интерес* и *Намерение*, которые соответствуют различным предметным областям, например *Спорт, Здоровье, Религия* и т.п.

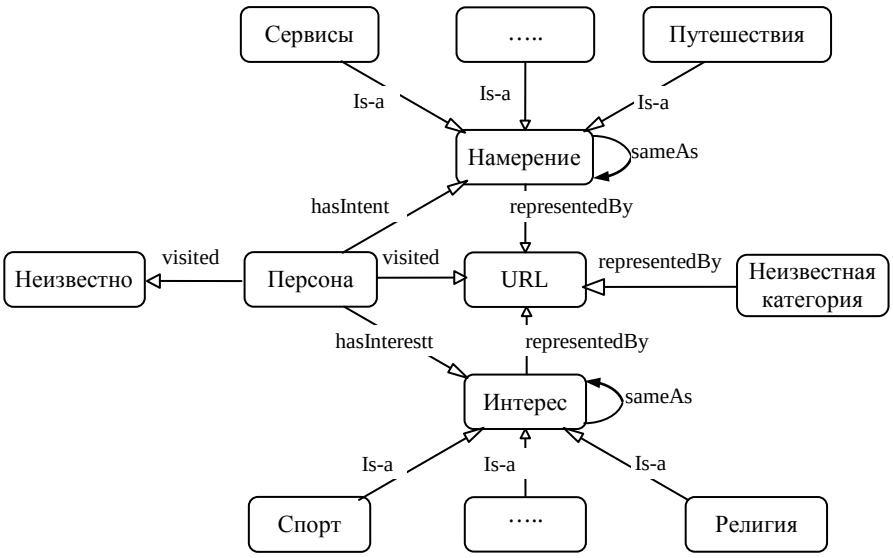

Рисунок 1 – Онтология, разработанная в [20] для представления профиля пользователя (для её графического представления использовано словесное описание, приведённое в указанной работе)

Основные отношения предложенной онтологии - это *hasInterest* и *hasIntent*. Отметим, что авторы разделяют все предпочтения пользователя на *Интересы* и *Намерения*. Намерения представляют сущности, которые существуют определённый период времени, в то время как *Интересы* полагаются постоянными для *Персоны*.

Отметим еще одну особенность данной работы. *OpenDNS* позволяет выполнять категоризацию только для доменов верхнего уровня и предоставляет информацию только об общей тематике сайта. Иначе говоря, если, например, пользователь разместил в твите ссылку на страницу своей любимой команды на сайте, посвященном футболу, то запрос к *OpenDNS* по данной ссылке, вернёт категории всего сайта, например, *Спорт*, *Футбол* и т.п. Но пользователь совсем необязательно интересуется футболом и спортом в целом. Возможно, его интересует только его любимая команда или какой-то из её игроков.

Таким образом, методика, предложенная в [20], позволят извлечь только достаточно общие интересы пользователя с невысокой, по нашему мнению, точностью. Тем не менее, подход может быть использован для первичного извлечения областей интересов пользователя, которые затем могут быть уточнены, например, путем анализа содержания страниц, размещённых по URL-адресу из твита. Описанная модель профиля пользователя также не предусматривает учёт контекста, в котором пользователь имеет тот или иной интерес. В планы авторов входит дальнейшее развитие технологии, например, анализ текстов, сопутствующих URL-адресу из твита, для получения лучших результатов.

Несмотря на недостатки, перечисленные выше, данная работа хорошо иллюстрирует, как можно *обогатить информацию о пользователе*, извлеченную из публичных источников, с помощью онтологий и фольксономий, разработанных сообществом, например, *OpenDNS* и *DBpedia*. Данная работа является еще одним подтверждением общего мнения о том, что использование таких дополнительных источников знаний является неотъемлемой частью рекомендующих систем третьего поколения, а также может быть использована для автоматизации построения онтологий.

Работа [27] предлагает словарь для описания профиля пользователя, представленный на языке *RDF*, что дает возможность его применения сразу в нескольких классах приложений, ориентированных на веб-поиск. Идея авторов состоит в том, чтобы использовать такой профиль пользователя в качестве ограничения совместно с лексикой запроса клиента («*уточне-*

ние запроса на основе интересов» в объяснении авторов). Авторы мотивируют такой подход тем, что обычно запрос бывает слишком расплывчатым, а потому без дополнительного уточнения он может приводить к генерации большого количества ненужных документов. В работе предлагается расширенное определение понятия *интерес пользователя*, которое включает в себя не только само понятие *интерес* («*тема, по которой агент хочет что-либо узнать, чему-либо обучиться или быть в нее вовлечен*», в соответствии с [27]), но и добавляет к нему некоторые атрибуты, задающие контекст, например, время. В реальности это расширение понятия *интерес* предполагает создание стандартного словаря понятий, представляющих интересы, связанные с фиксированным набором атрибутов, а также с некоторыми контекстными атрибутами. Но, во-первых, в настоящее время не представляется возможным разработать словарь, охватывающий все виды интересов, во-вторых, работа игнорирует такую важную компоненту спецификации профиля пользователя, как структура интересов пользователей. В сущности, результаты этой работы применимы, главным образом, к веб-поиску.

Статья [28] мотивирует использование онтологии необходимостью восполнить нехватку семантической информации при построении персонализированного профиля пользователя, который может динамически меняться с течением времени. Как известно, рекомендации, вырабатываемые на основе коллаборативной фильтрации, основаны на *семантическом* сходстве целевого пользователя с другими пользователями и на мнениях последних (они представляются значениями рейтингов) относительно рекомендуемого экземпляра товара/услуги. Основываясь на онтологии интересов пользователя, работа предлагает модель профиля пользователя и алгоритм, выполняющий ее автоматическое обновление по мере получения новых данных. Иерархическая структуризация интересов формулируется только в качестве цели будущих исследований. Поэтому результаты этой работы не могут напрямую использоваться в рекомендующих системах третьего поколения.

Резюмируя содержание работ, рассмотренных в данном разделе, а также других, которые здесь не упоминаются, можно сделать следующее заключение. Эти работы, хотя и подготовили необходимый теоретический и технологический базис для широкого использования онтологических моделей в качестве единой структуры для представления различных компонент знаний рекомендующих систем, тем не менее, нуждаются в дополнительных исследованиях. Особенно убедительным этот тезис представляется применительно к специфическим видам информации, которую необходимо использовать в рекомендующих системах третьего поколения. Некоторые из них рассматриваются в последующих разделах.

5 Категоризация тэгов для построения персонифицированного профиля пользователя

Тэгами называют семантически осмысленные метки, которые любой пользователь может присваивать некоторым Интернет-ресурсам, например, документам, видео и аудио файлам, информации, представленной в социальных сетях и т.п. Иногда они играют роль, подобную ключевым словам, иногда они выражают эмоциональное отношение пользователя к соответствующему ресурсу, иногда используются для улучшения поиска. В любом случае тэги содержат весьма полезную информацию для идентификации свойств Интернет-ресурсов и отношения к ним пользователя. Роль тэгов многообразна, однако в данной работе они рассматриваются с позиций их полезности для извлечения информации об интересах пользователя.

Работы, посвященные использованию тэгов для построения онтологии предметной области рекомендаций и профиля пользователя, появились недавно. К настоящему времени эта проблема все еще разработана недостаточно, хотя она исследуется уже около 8 лет. Рассмотрим некоторые работы, которые посвящены решению указанной проблемы.

В работе [29] предлагается методология моделирования онтологического профиля пользователя с использованием тэгов Интернет-ресурсов и сервисов, к которым пользователь проявляет интерес. Использование информации о тэгах позволяет вовлекать в построение онтологического профиля пользователя дополнительные знания. Заметим, что информация о тэгах часто носит персонифицированный характер и потому важна для построения персонального профиля пользователя.

В качестве примера в статье [29] рассматриваются сервисы *Flickr* и *Delicious*. Для сведения, *Flickr* представляет собой сервис, предназначенный для хранения и дальнейшего использования цифровых фотографий и видеороликов. *Delicious* [30] – это веб-сайт, предоставляющий зарегистрированным пользователям бесплатную услугу по хранению и публикации *закладок* на страницы Интернет. Поскольку информация об интересах пользователя распределена по очень многим источникам в Интернет, то дополнительной мотивацией работы [29] является отработка технологии построения онтологии по информации из различных источников. В частности, авторы говорят о технологии консолидации тэгов, распределенных по многим источникам, в рамках онтологии.

Информация о профиле пользователя, сформированная на основе множества тэгов, удобна для реализации механизма сравнения (англ. *matching*) ее с семантической аннотацией рекомендуемого продукта/услуги, представленной в терминах *той же онтологии,* построенной для того же множества тэгов. Такое семантическое сходство рассматривается в работе как основа персонификации профиля пользователя и, соответственно, персонификации рекомендаций.

Авторы предлагают *отображать* теги на понятия онтологии профиля пользователя с помощью словарей *WordNet* и категорий Википедии. Напомним, что *WordNet* – это электронный тезаурус/семантическая сеть для английского языка, разработанный в университете Принстон (США) и выпущенный вместе с сопутствующим программным обеспечением под свободной лицензией [31]. Отображение тэгов на понятия профиля пользователя реализуется процедурой из трех шагов: (а) фильтрация тегов, (б) получение семантической информации о тегах из сети Интернет и (в) категоризация полученных понятий согласно классам понятий существующей онтологии.

Опишем несколько подробнее процесс извлечения и обработки тегов пользователя, а также процесс их отображения на понятия онтологии с использованием семантической информации, полученной из сети Интернет.

На первом шаге, на котором выполняется фильтрация тэгов, сначала выбирается некоторый общий английский словарь терминов, например, *WordNet* [31-32], Википедия и/или Google и выполняется предобработка тегов с использованием их морфологических и семантических трансформаций с последующим связыванием результирующего множества тэгов с терминами выбранного словаря. Лексический фильтр отбрасывает слишком короткие тэги (длиной в один символ), слишком длинные (более 25 символов), а также артикли, предлоги и т.п. Выполняется также преобразование специальных символов (например, символов *à, á, â* в символ *a*). Редко встречающиеся теги также удаляются с помощью пороговой фильтрации, хотя последнее и не является однозначно полезным. Оставшееся множество тэгов передается в *WordNet,* после чего выполняется их сравнение (англ. *matching*) с терминами словаря (по точному их совпадению). Далее «успешные тэги» (те, для которых нашлась пара в выбранном словаре) добавляются в результирующий набор тэгов. Для уточнения правильности написания тэгов используется также известный механизм Google «*возможно, вы имели в виду*». С помощью алгоритма, разработанного авторами [29], составные существительные типа *unitedkingdomsouthampton, san_francisco* и т.п. разбиваются на части. Полученные теги снова проверяются с помощью словаря *WordNet*. Теги, которые не были идентифицированы с по-

мощью *WordNet* (имена собственные, сокращения или сленг), далее анализируются на предмет поиска «пары» с помощью дампа категорий Википедии. Часто используемые аббревиатуры заменяются общепринятыми терминами, например, *nyc* заменяется на *New York City*. Используется также морфологический анализ, который позволяет удалить морфологически одинаковые теги, например, *blog, blogs, blogging*. Затем, с помощью алгоритма, снова использующего *WordNet*, удаляются синонимы.

На втором шаге из Википедии извлекается семантическая информация по отфильтрованным тегам. Эта информация включает в себя наименование понятия, соответствующего тегу, принятое в Википедии, и его категорию.

Наконец, *на третьем шаге* выполняется отображение понятий из Википедии, которые соответствуют тегам, на понятия разработанной ранее онтологии. Тем самым выполняется расширение ранее построенной онтологии. Для этого выполняется морфологическое сравнение наименований понятий из Википедии и наименований его категорий с классами разработанной ранее онтологии. Понятие, сформированное на основе тэгов, добавляется в онтологию как экземпляр класса с наиболее похожим именем (категории, извлеченные из Википедии, добавляются к экземпляру как *RDFS*-метки).

Для формирования профиля пользователя авторы [29] используют представление предпочтений пользователя в следующей форме [33–34][6]:

$$U_m = [u_{m,1},...,u_{m,i},...,u_{m,K}]^T \text{ , где } u_{m,i} \in [0,1],\ c_m \in (c_1,...,c_M)\ .$$

Бинарные компоненты $u_{m,i}$ вектора U_m выражают меру интереса пользователя $c_m \in C$ к понятию o_i онтологии предметной области O или отдельных экземпляров этого понятия, т.е. $o_i \in O$; K – общее число понятий в онтологии. Рекомендуемые продукты/услуги (*items*) $s_n \in \boldsymbol{S}$ описываются аналогичным вектором:

$$V_n = [v_{n,1}, v_{n,2},...,v_{n,i}..,v_{n,K}]^T \text{ , где } v_{n,i} \in [0,1],\ s_n \in (s_1,...,s_N)\ ,$$

в котором бинарные компоненты $v_{n,i}$ вектора V_n выражают вес понятия $o_i \in O$ для товара/услуги $s_n \in (s_1,...,s_N)$, K – общее число понятий в онтологии.

Далее, в этом подходе вместо матрицы рейтингов $\boldsymbol{R} = \{r_{i,j}\}|_{i=1,\ j=1}^{M,\ N}$, введенной в разделе 1, используется матрица $\boldsymbol{D} = \{d_{m,n}\}|_{i=1,\ j=1}^{M,\ N}$, в которой элемент $d_{m,\ n} \in \boldsymbol{D}$ равен косинусу угла между векторами $U_m = [u_{m,1},...,u_{m,i},...,u_{m,K}]^T$ и $V_n = [v_{n,1}, v_{n,2},...,v_{n,i}..,v_{n,K}]^T$. Эта величина используется в качестве меры сходства вектора интересов U_m пользователя $c_m \in C$ и вектора V_n , описывающего свойства товара/услуги $s_n \in S$. Напомним, что значение косинуса между парой векторов равно их нормированному скалярному произведению:

$$d_{m,n} = (U_m\ \ ,V_n\ \ )/(\|U_m\| \times \|V_n\|).$$

При выработке рекомендации для пользователя $c_m \in C$ предпочтение отдается товару/услуге $s_n \in S$ с наименьшим значением меры $d_{m,n}$.

Очевидный недостаток описанного подхода к построению персонифицированного профиля пользователя с использованием информации о тэгах состоит в том, что размерность бинарных векторов $U_m \in \boldsymbol{U}$ и $V_n \in \boldsymbol{V}$ всегда равна числу понятий в онтологии предметной области, а оно может исчисляться тысячами, и их число может увеличиваться в процессе работы системы. Матрица «*пользователь–продукт*» будет очень большой размерности. Работа

[6] Здесь и далее используются обозначения, частично введенные ранее в разделе 1.

с такой моделью потребует огромных вычислительных ресурсов. Однако, поскольку здесь вычисления ведутся с бинарными переменными, то не возникнет эффекта накопления ошибок, свойственного процедурам обработки больших данных с вещественными атрибутами.

Среди преимуществ этого подхода авторы выделяют инвариантность технологии его реализации к предметной области и возможность ее применения для любых продуктов и услуг.

Описанный подход и реализующая его технологии проверены с помощью системы рекомендации новостей *News@hand*, разработанной авторами реферируемой статьи. В этой системе поддерживается также автоматическое получение новостных статей из сети, автоматическое семантическое аннотирование этих статей с использованием средств обработки естественных языков [35] и средства *Lucene* [36].

В целом, работа достаточно убедительно демонстрирует интересную и практически реализуемую перспективу комбинированного использования *фольксономии* (*folksonomy*)[7], как иначе принято называть словари и базы данных тэгов и знаний, которые могут быть получены из различных источников с использованием иерархической категоризации понятий (онтологической информации), содержащейся в словарях, в частности, в дампе категорий Википедии.

Интересный подход использования тэгов для построения онтологий предложен в работе [37]. Он, хотя и не имеет прямого отношения к построению онтологического профиля пользователя в рекомендующих системах, тем не менее, предлагает новый подход к построению онтологий с использованием тэгов, который представляется достаточно продуктивным и для рекомендующих систем, в частности, для построения онтологического профиля пользователя на основе тэгов. С методологической точки зрения, все основные операции этого алгоритма направлены на максимальное повторное использование имеющегося опыта разработок, доступного авторам.

На первом шаге построения онтологии предложенный алгоритм генерирует некоторую базовую (черновую) онтологию, которая содержит в себе множество тэгов данных, каждому их которых поставлен в соответствие термин словаря (понятие). Они структурированы некоторым множеством отношений. Для автоматической генерации базовой онтологии используется распределенная структура фольксономии, представленная в ресурсе *OpenDNS* [21]. Этот ресурс, созданный волонтерами с помощью накопления и объединения собственных фольксономий (результатов по аннотации тэгов), представляет собой достаточно простую (в первоисточнике англ. - *lightweight*) концептуальную структуру, которая связывает пользовательские термины с веб-ресурсами тэгов. Эта процедура позволяет существенно ускорить концептуализацию и классификацию тэгов данных и, тем самым, ускорить построение онтологии. Понятия и отношения базовой онтологии, построенной таким способом, далее корректируются экспертом по предметным знаниям. В этом процессе какие-то понятия и отношения из базовой онтологии удаляются, какие-то могут добавляться.

Второй шаг описываемого подхода также опирается на имеющийся опыт построения онтологий и пытается максимально его использовать. На этом шаге авторы обращаются к облачному сервису *веба данных* (англ. *Linked Data*) [19] для того, чтобы использовать внешние связи к классам понятий этого ресурса и к отношениям. Эта работа выполняется предметными экспертами. Затем выполняется совместный просмотр аннотаций и отношений веб сайтов веба данных, которые были рекомендованы экспертами, и сайтов фольксономии, которые были использованы при построении базовой онтологии.

[7] Фольксономии обычно содержат термины, которые отсутствуют в стандартных глоссариях.

Исследования в области методов концептуализации тэгов и построения онтологий на этой основе ведутся в настоящее время достаточно активно (см., например, [38]). Расширенный обзор по состоянию исследований в этой области можно найти в работе [39].

6 Построение профиля пользователя в контексте социальных сетей: пример

Широкий и важный класс задач выработки рекомендаций охватывают приложения, функционирующие в контексте социальных сетей. Одним из множества имеющихся примеров такого класса приложений является рекомендующая система поиска экспертов для ответа на вопросы, рассмотренная в работе [40], когда эксперты образуют некоторую сеть. В этом приложении эксперты выступают в роли «пользователей», которым адресуется («рекомендуется») некоторый запрос (продукт, услуга – в терминологии традиционной постановки задачи), и задача состоит в том, чтобы адресовать этот запрос экспертам, которые наиболее компетентны по тематике запроса. Компетентность эксперта здесь играет роль, аналогичную роли персонального профиля пользователя. Другой пример – задача подбора специалистов на некоторые должности при заданных требованиях к профессиональным знаниям для каждой из должностей. Главной задачей здесь, по-прежнему, является задача построения *персонального профиля* пользователя, которым является *эксперт*, в рассматриваемом приложении. Особенность этой задачи состоит в том, что при поиске рекомендаций используется «сетевой контекст», задаваемый топологией и числовыми характеристиками социальной сети экспертов. Этот сетевой контекст задает меры сходства различных пользователей (экспертов – в задаче [40]) и связи между ними, а значит, позволяет использовать методы коллаборативной фильтрации при формировании рекомендаций.

Рассмотрим пример сетевого контекста для приложения из [40]. В социальной сети экспертов каждый из них имеет некоторую область экспертизы, описываемую множеством ключевых слов. Сходство областей экспертизы любой пары экспертов (узлов социальной сети, в общем случае), которое задается формально некоторой моделью, определяет то, что названо выше «сетевым контекстом». Заметим, что постановка задачи, рассматриваемая в работе [40], обобщается на широкий класс приложений, решающих задачи формирования рекомендаций в социальных сетях, например, поиск друзей в сети, формирование групп по некоторому множеству интересов и прочие.

Опишем кратко, каким образом сетевой контекст (сходство на множестве пар экспертов) описывается и вычисляется формально в задаче, рассматриваемой в [40]. Эта модель представляет определенный интерес, поскольку является одним из примеров меры *семантического сходства пользователей*, столь важной для применения методов коллаборативной фильтрации.

Сеть экспертов (в общем случае – социальная сеть) описывается графом, в котором каждому эксперту сети ставится в соответствие узел графа, а в качестве атрибутов узла рассматриваются ключевые слова, задающие область экспертизы (область знаний) эксперта, представленного узлом. Дуги между вершинами этого графа задают отношение цитирования одного эксперта другим, так что это отношение не является симметричным. Эти данные используются для вычисления матрицы сходства на множестве узлов графа. Авторы используют совместно несколько мер, которые затем некоторым образом комбинируются в итоговой мере сходства. Одна из этих мер называется *структурной* мерой сходства. Она задается либо числом общих узлов, связанных описанным выше отношением цитирования с каждым элементом пары (в общем случае – числом общих соседей рассматриваемой пары узлов графа), либо числом других авторов, которые цитируют оба элемента пары. В общем случае вид меры сходства зависит от приложения.

Что касается модели персонального профиля пользователя, то работа [40] использует для этих целей модель, которая для каждого пользователя является экземпляром некоторого класса предметной онтологии. Другими словами, область экспертизы каждого эксперта представляется множеством терминов. Для этих терминов с использованием английского словаря терминов *WorldNet* отыскиваются наиболее близкие понятия предметной онтологии, причем близость оценивается с помощью весов релевантности (к сожалению, авторы не сообщают, каким именно образом эти веса вычисляются). Профиль пользователя определяется тем классом понятий предметной онтологии, который наиболее близок к представлению профиля пользователя.

По-видимому, на момент написания статьи авторы [40] еще не закончили разработку модели онтологического профиля пользователя, поскольку в работе представлена только модель концептуального уровня. В целом, предложенный подход к построению онтологического профиля пользователя проработан достаточно слабо, поэтому о новых решениях в этой части авторы не сообщают.

7 Онтологические модели профиля пользователя для множества источников данных

Необходимость формирования профиля пользователя по данным, распределенным по множеству источников, возникает практически в каждом случае, если не считать те случаи, которые задаются стандартными наборами данных, предназначенными для проверки и сравнения различных моделей построения персонального профиля пользователя (англ. *benchmarks*). По этой причине необходимо уметь работать с распределенными источниками данных, в которых проявляются «следы» пользователя. Обычно для решения таких задач привлекается опыт исследователей в области слияния данных из различных источников (англ. *data fusion*). Этот опыт достаточно большой, он дает много примеров использования онтологий, и поэтому здесь об аналогичных задачах применительно к построению онтологического профиля пользователя говорится только с той целью, чтобы обозначить существование такой проблемы в области рекомендующих систем.

Одна из таких моделей предлагается в работе [41]. В ней предполагается, что для построения профиля пользователя привлекается информация о его работе в социальных сетях, например, в *Twitter, Facebook, LinkedIn*, а также используются данные с его домашней страницы.

Авторы рассматривают разные источники данных как результаты, измеряемые разными сенсорами, а саму задачу объединения информации – как задачу слияния сенсорной информации (англ. *multi-sensor information fusion*). Рассматриваются три стратегии слияния информации, которые традиционно приняты в теории слияния данных и информации, а именно [42]:

1) *Слияние информации на уровне данных*, когда данные из разных источников собираются в общей базе и обрабатываются централизованно. Обычно это возможно при небольших размерностях данных о пользователе. Здесь, однако, затруднительно обеспечить конфиденциальность данных, которые в разных источниках могут иметь разных владельцев.
2) *Слияние информации на уровне признаков, атрибутов данных* (англ. *feature level fusion*), когда в каждом источнике данных формируется потенциальное множество интересов, они объединяются вместе в общей онтологии, и далее производится их фильтрация по важности и персонификация. Этот вариант слияния данных несколько лучше предыдущего, однако, страдает многими аналогичными недостатками.

3) *Слияние информации на уровне решений.* В этом случае информация об интересах и предпочтениях пользователя, представленная в терминах онтологий, формируется в каждом источнике независимо друг от друга. Затем понятия и отношения локальных онтологий объединяются с помощь метаонтологии, для которой понятия локальных онтологий являются подклассами ее классов. Этот подход, в общем случае, представляется наиболее перспективным. Он позволяет успешно преодолевать различные проблемы, связанные и с большим масштабом объединенных данных, и с конфиденциальностью. Кроме того, в каждом источнике данных может использоваться своя стратегия разработки онтологии интересов пользователя. Авторы работы [41] принимают именно эту стратегию.

Интересы пользователей описываются в ней в терминах понятий и отношений онтологии. Онтология используется для установления различных взаимосвязей между интересами пользователя, представленными в нескольких источниках данных с различными степенями детализации. Онтология верхнего уровня, объединяющая несколько источников данных, используется также для вывода неявных интересов пользователей путем рассуждений с использованием иерархии интересов в онтологии. В целом, в работе кратко описываются потенциальные направления и перспективы богатой семантически-ориентированной модели профиля пользователя, но не более того, так как основные идеи демонстрируются в работе на слишком упрощенных примерах. Пока не ясно, насколько адекватно предлагаемые идеи могут быть использованы в приложениях реальных масштабов.

Заключение

В рекомендующих системах третьего поколения онтологии рассматриваются в качестве базовой семантической модели и структуры для представления персонального профиля пользователя. В ней модель знаний о профиле пользователя представляется в терминах классов понятий, которые описывают онтологию предметной области рекомендаций. Сформулируем основные выводы о роли онтологий в рекомендующих системах третьего поколения, в частности, о представлении персонального профиля пользователя, а также о некоторых особенностях такого представления.

1) Онтология позволяет представлять интересы пользователя в терминах семантически интерпретируемых категорий естественного языка, т.е. в терминах, понятных пользователю, в которых он о своих интересах обычно говорит. Например, интересы формулируются в таких терминах, как интерес к фильмам комедийного жанра или интерес пользователя к фильмам с участием конкретных актеров, интерес к новостям на конкретную тему и т.п. Семантически понятная интерпретация интересов пользователя позволяет получать обратную связь от пользователя как при построении онтологии и профиля пользователя, так и при оценке качества работы системы в целом. В онтологической модели профиль пользователя представляется как структурированное подмножество его интересов, которые представлены конкретными понятиями/категориями предметной области, или/и экземплярами некоторых классов понятий. Такая семантически ясная модель профиля пользователя может дать полезную информацию для объяснения рекомендаций. Особенно полезной эта информация является в том случае, когда интересы пользователя построены на основе поиска причинных связей интересов и оценок, которые он дает продуктам/услугам.
2) При построении профиля пользователя разработчики должны стремиться использовать всю доступную информацию, в которой, так или иначе, проявляются «следы» поведения пользователя, отражающие его позитивные или негативные интересы. В особенности полезной является информация, в которой представлены субъективные оценки пользовате-

ля, его отношение к тем или иным продуктам/услугам. Одним из важных источников такой семантически-интерпретируемой информации является множество тэгов. Ими пользователи метят данные. Часто пользователи отражают в тэгах информацию, которую невозможно или трудно описать в терминах свойств товаров/услуг. Поэтому иногда для оценки мнения пользователя тэги могут быть полезнее, чем любая другая информация. Они чаще являются *субъективными* оценками и потому дают важную информацию для персонификации рекомендаций.

К настоящему времени проблема использования тэгов для формирования знаний о профиле пользователя пока исследована недостаточно, хотя она, очевидно, и является проблемой первостепенной важности. Сейчас эта область исследований и разработок остается все еще на уровне отдельных удачных решений, но не на уровне хорошо отработанной технологии. Она требует серьезных дополнительных исследований, которые должны быть направлены, прежде всего, на создание *автоматизированных* технологий обработки тэгов для построения онтологического профиля пользователя. Некоторые примеры таких технологий, а также методы и средства их автоматизации были рассмотрены в данном обзоре.

3) Большой интерес представляют исследования, которые для построения онтологии и профиля пользователя используют не только ранее накопленную (статическую) информацию, но и динамически отслеживают новые интересы пользователя и трансформируют его профиль на основе новых данных. Очевидно, что методы обучения профиля пользователя должны быть инкрементными, т.е. они должны уточнять модель профиля пользователя по мере поступления новой информации. Во многих случаях это уточнение сводится к расширению множества примеров уже существующих классов понятий и модификации правил выработки рекомендаций.
4) В рекомендующих системах третьего поколения необходимо решать проблему кросс-доменных рекомендаций. Онтологическая модель профиля пользователя позволяет существенно упростить решение этой задачи, если его представить единой структурой на множестве понятий-интересов пользователя, которые относятся к различным предметным областям. Профиль пользователя при этом не должен быть представлен просто множеством локальных подмоделей, каждая из которых относится к своей предметной области. Их объединение с помощью понятий метауровня и использование технологий «выравнивания онтологий» (англ. *ontology alignment*) и другие приемы современной инженерии онтологий позволят связать знания из разных областей в единую модель.
5) Во многих работах онтология строится так, чтобы любому ее понятию можно было бы поставить в соответствие множество конкретных примеров из базы данных. Такая структура является предметом исследований в области анализа формальных понятий (англ. *Formal Concept Analysis*) [43]. Эта опция применения онтологий представляется весьма полезной: она позволяет проводить динамическую переоценку важности тех или иных интересов для пользователя по мере получения новых данных. Если рекомендующая система функционирует на продолжительном интервале времени (измеряемом годами), то информация о примерах тех или иных понятий онтологии и об их динамике позволит отслеживать «возрастную» динамику интересов пользователя, включая динамику его интересов в различных контекстах. Другие достоинства такой структуры представления обучающих данных кратко описаны в [44].
6) В данном обзоре не рассмотрены вопросы использования онтологий для представления зависимостей профиля пользователя от контекста, хотя именно эта компонента профиля пользователя во многом определяет современные тенденции развития теории и практики рекомендующих систем. Контекстно-зависимые рекомендации и роль онтологий в таких

системах – это очень обширная и важная тема, которая достойна отдельного обзора. Учёт контекста в рекомендующих системах исследуется уже около 12–15 лет. В настоящее время достигнутые в ней результаты позволяют создавать достаточно эффективные рекомендующие системы. Тем не менее, некоторые важные аспекты этой проблемы исследованы недостаточно. Например, пока нет однозначного мнения о том, как отделить контекст от самой предметной области рекомендаций и нужно ли это делать вообще, если принимать во внимание тенденцию использования онтологической модели контекста, интегрированной в предметную онтологию.

Среди ключевых проблем контекстно-зависимых рекомендующих систем авторы работы [8] упоминают необходимость сравнительного изучения разных моделей встраивания контекста в технологию поиска рекомендаций. Пока неясно, какая из предложенных моделей, а именно предварительная фильтрация, пост-фильтрации и встраивание контекста в модель знаний предпочтительнее и в каких случаях. По-видимому, возможны также и альтернативные методы учета контекста. Не менее важным является вопрос о полезности их разумной комбинации. Требует дополнительного исследования задача категоризации контекстов и возможности автоматизации построении онтологии контекстов. Необходимо более глубоко изучить корректность использования процедур обобщения контекстов, а также построить различные, например, топологические метрики оценки близости контекстов. Это поможет справиться с проблемой дефицита данных, которая характерна для задач обучения профиля пользователя в контекстно-зависимых системах.

Следует обратить внимание на то, что все проблемы вычислительной эффективности и качества решений, которые возникают в области интеллектуальной обработки *больших данных*, во многом характерны и для приложений в области рекомендующих систем третьего поколения. В обоих случаях необходимо решать задачу обработки данных большого объема и размерности. Как и в задачах построения рекомендующих систем, в задачах из области больших данных остро стоит проблема автоматизации процессов построения онтологий: в них число базовых понятий может исчисляться тысячами и более. Это говорит о том, что по сути своей задачи в области рекомендующих систем третьего поколения и задачи в области обработки больших данных во многом схожи. Поэтому исследования в этих актуальных областях могут и должны обогащать друг друга.

Благодарности

Данная работа выполнена при частичной поддержке Исследовательского Центра Самсунг, г. Москва, а также проекта 1.12 Программы фундаментальных исследований отделения нано- и информационных технологий Российской академии наук «Интеллектуальные информационные технологии, системный анализ и автоматизация».

Список источников

[1] ***Tuzhilin, A.*** Keynote presentation at International Conference on Data Mining (ICDM 2013) / A. Tuzhilin. - Dallas, Texas, December, 2012.

[2] ***Cao, L.*** Actionable knowledge discovery and delivery / L. Cao // In: WIREs Data Mining and Knowledge Discovery, Volume 2, March/April 2012, John Willey & Sons , Inc., 2012. – P. 149–163.

[3] ***Argyri, C.*** Actionable Knowledge / C. Argyri // The Oxford Handbook of Organization Theory. Edited by Knudsen C. and Tsoukas H. - Oxford University Press, 2005.

[4] Recommender Systems Handbook / Ricci F., Rokach L. and Shapira B (Eds.). - Springer, 2011. – 842 P.

[5] ***Jannach, D.*** Tutorial: Recommender Systems / D. Jannach, G. Friedrich // International Joint Conference on Artificial Intelligence.(Beijing, August 4, 2013). Available at http://ijcai-11.iiia.csic.es/files/proceedings/ Tutorial%20IJCAI%202011%20Gesamt.pdf (Актуально на 18.07.2014).

[6] ***Segaran, T.*** Programming Collective Intelligence / T. Segaran // O'Reilly, 2006 (Русский перевод: Тоби Сегаран. Программируем коллективный разум. Издательство Символ +, 2008. – 368 с.
[7] ***Adomavicius, G.*** Incorporating contextual information in recommender systems using a multidimensional approach / G. Adomavicius, R. Sankaranarayanan, S. Sen, A. Tuzhilin // ACM Transactions on Information Systems (TOIS). – 2005 - 23(1). - P. 103–145.
[8] ***Adomavicius, G.*** Context-Aware Recommender Systems / G. Adomavicius, B. Mobasher, F. Ricci, A. Tuzhilin // AI Magazine, FALL 2011. - P. 67–80.
[9] ***Adomavicius, G.*** Context-Aware Recommender Systems / G. Adomavicius, A. Tuzhilin // In Ricci F., Rokach L, Shapira B,·Kantor P (Eds.). Recommender Systems Handbook, Springer, 2011. - P. 217–256.
[10] ***Gauch, S.*** Ontology-Based Personalized Search and Browsing / S. Gauch, J. Chaffee, A. Pretschner // ACM Web Intelligence and Agent System. - 2003. - Vol. 1. - No. 3/4. - P. 219–234.
[11] ***Leung, K.W.T.*** Deriving Concept-Based User Profiles from Search Engine Logs / K.W.T. Leung, D.L. Lee // IEEE Transaction on Data and Knowledge Engineering. - 2010. - Vol. 22. - No. 7. - P. 969–982.
[12] ***Liu, F.*** Personalized Web Search by Mapping User Queries to Categories / F. Liu, C. Yu, W. Meng // In Proc. of Intern. Conf. on Information and Knowledge Management (CIKM), 2002.
[13] ***Xu, Y.*** Privacy-Enhancing Personalized Web Search / Y. Xu, K. Wang, B. Zhang, Z. Chen // Proceedings of World Wide Web (WWW) Conference, 2007. - P. 591–600.
[14] Open Directory Project, http://www.dmoz.org/ (Актуально на 18.07.2014).
[15] ***Costa, A.C.*** Cores: Context-aware, ontology-based recommender system for service recommendation / A.C. Costa, R.S.S. Guizzardi, J.G.P. Filho // In Proc. 19-th Intern. Conf. on Advanced Information Systems Engineering (CAISE07). 2007.
[16] ***Middleton, S.E.*** Ontological user profiling in recommender systems / S.E. Middleton, N.R. Shadbolt, D.C. de Roure // ACM Transaction on Information Systems. - 2004. - 22(1). - P. 54-88.
[17] ***Trajkova, J.*** Improving Ontology-Based User Profile / J. Trajkova, S. Gauch // RIAO, 2004. - P. 380-390.
[18] ***O'Hara, K.*** The AKT Manifesto / K. O'Hara, N. Shadbolt, S. Buckingham. - http://citeseer.uark.edu:8080/citeseerx/ showciting?cid=3895624 (Актуально на 18.07.2014).
[19] ***Bizer, C.*** Linked Data – The Story So Far / C. Bizer, T. Heath, T. Berners-Lee // International Journal on Semantic Web and Information Systems (IJSWIS), 2009.
[20] ***Peña, P.*** Collective Knowledge Ontology User Profiling for Twitter / P. Peña, R. del Hoyo, J. Vea-Murguía , C. González, S. Mayo // 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.
[21] OpenDNS cloud websites tagging: http://community.opendns.com/domaintagging/ (Актуально на 18.07.2014).
[22] DBpedia–a crowd-sourced community effort to extract structured information from Wikipedia: http://dbpedia.org (Актуально на 18.07.2014).
[23] The Apache Cassandra database: http://cassandra.apache.org/ (Актуально на 18.07.2014).
[24] Virtuoso, a grade multi-model data server: http://virtuoso.openlinksw.com/ (Актуально на 18.07.2014).
[25] NoSQL data bases: http://ru.wikipedia.org/wiki/NoSQL (Актуально на 18.07.2014).
[26] Pellet reasoning server: http://clarkparsia.com/pellet/ (Актуально на 18.07.2014).
[27] ***Zeng, Y.*** User Interests: Definition, Vocabulary, and Utilization in Unifying Search and Reasoning / Y. Zeng, Y. Wang, Z.S. Huang, D. Damljanovic, Zh. Ning, C. Wang // In An A. et al. (Eds.): Active Media Technology 2010, Lecture Notes in Computer Science, vol. 6335, Springer, 2010. - P. 98–107.
[28] ***Su, Z.G.*** Research on Personalized Recommendation Algorithm Based on Ontological User Interest Model / Z.G. Su, J. Yan, H.F. Ling, H.S. Chen // J. of Computational Information Systems. – 2012. - Vol. 8. - No 1. - P. 169–181. Available also at http://www.Jofcis.com/ (Актуально на 10.07.2014).
[29] ***Cantador, I.*** Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations / I. Cantador, M. Szomszor, H. Alani, M. Fernández , P. Castells // In Proc of 1st Intern. Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008), Tenerife, Spain. 2008.
[30] Delicious: Social Bookmark manager: http://ru.wikipedia.org/wiki/Delicious (Актуально на 18.07.2014).
[31] WordNet: http://ru.wikipedia.org/wiki/WordNet (Актуально на 18.07.2014).
[32] ***Miller, G.A.*** WordNet: A Lexical Database for English / G.A. Miller // Communications of the Association for Computing Machinery. – 1995. - 38(11). - P. 39-41.
[33] ***Castells, P.*** An Adaptation of the Vector-Space Model for Ontology-based Information Retrieval / P. Castells, M. Fernández, D. Valle // IEEE Transactions on Knowledge and Data Engineering. – 2007. - 19 (2). - P. 261-272.
[34] ***Vallet, D.*** Personalized Content Retrieval in Context Using Ontological Knowledge / D. Vallet, P. Castells, M. Fernández, P. Mylonas, Y. Avrithis //IEEE Trans. on Circuits and Systems for Video Technology. – 2007. - 17(3). - P. 336-346.

[35] ***Alfonseca, E.*** The Wraetlic NLP Suite / E. Alfonseca, A. Moreno-Sandoval, J.M. Guirao, M. Ruiz-Casado // In Proc. of the 5th Intern. Conf. on Language Resources and Evaluation. 2006. - P. 2277–2280. Available at http://www.lrec-conf.org/proceedings/lrec2006/ (Актуально на 18.07.2014).
[36] Lucene: An Open Source Information Retrieval Library: http://lucene.apache.org/ (Актуально на 18.07.2014).
[37] ***García-Silva, A.*** Social Tags and Linked Data for Ontology Development: A Case Study in the Financial Domain / A. García-Silva, L.J. García-Castro, A. García // In Proceedings of 4–th International Conference on Web Intelligence, Mining and Semantics (June 2–4, 2014, Thessaloniki, Greece).
[38] ***Jaschke, R.*** Discovering shared conceptualizations in folksonomies / R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, G. Stumme// Web Semantics Science Services and Agents on the World Wide Web. - 2008. - 6(1). - P. 38-53.
[39] ***Garcia-Silva, A.*** Review of the state of the art: discovering and associating semantics to tags in folksonomies / A. Garcia-Silva, O. Corcho, H. Alani, A. Gomez-Perez // The Knowledge Engineering Review. - 2012. - 27(2). - P 57-85.
[40] ***Kadima, H.*** Toward ontology-based personalization of a Recommender System in social network / H. Kadima, M. Malek // International Journal of Computer Information Systems and Industrial Management Applications. - 2013. - Vol. 5. - P. 499-508.
[41] ***Ma, Y.F.*** User Interests Modeling Based on Multi-source Personal Information Fusion and Semantic Reasoning / Y.F. Ma, Y. Zeng, R. Xu, Zh Ning // In Zhong N., Callaghan V., Ghorbani A., Hu B. (Eds.): Active Media Technology-2011, Lecture Notes in Computer Science, vol. 6890, Springer, 2011. - P. 195–205.
[42] ***Varshney, P.K.*** Multisensor data fusion / P.K. Varshney // Electronics & Communication Engineering J. – 1997. - 9(6). - P. 245–253.
[43] ***Ganter, B.*** Formal concept analysis: foundations and applications / B. Ganter, R. Wille. - Springer, 1999.
[44] ***Gorodetsky, V.*** Agent-based Customer Profile Learning in 3G Recommending Systems / V. Gorodetsky, V. Samoylov, O. Tushkanova // In: Proc. of the 9-th Intern. Workshop "Agent and Data Mining Interaction" (ADMI - 2014) associated with International Conference "Autonomous Agents and Multi-agent Systems" (AAMAS - 2014) (Paris, May 5–9 2014). To be also published in Post Proceedings of the Workshop as a volume of Lecture Notes in Computer Science, Springer, 2014.

ONTOLOGY-BASED USER PROFILE PERSONIFICATION IN 3G RECOMMENDER SYSTEMS

V.I. Gorodetsky[1], O.N. Tushkanova[2]

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
[1]gor@iias.spb.su, [2]tushkanova@iias.spb.su

Abstract

3G recommender system (3G RecSys) is a novel and, in basic aspects, future paradigm of human decision support/prediction systems operating based on human-like semantic categories in knowledge representation, knowledge discovery and knowledge usage. According to the experts' opinion, it is expected that 3G RecSys will focus on semantically transparent personal user profile structuring his/her multidimensional personal interests and preferences. It will bring novel perspectives from data mining and knowledge discovery, while emphasizing user's decision explanation and detection of context-aware causality core determining this or that user's choice. As an effect of these capabilities, 3G RecSys should be capable, in what concerns with recommendation proposed by it, to answer "why?" questions using explanatory user interface. At present days, ontology is recognized as the most natural well-developed modeling framework and knowledge representation paradigm aimed at formal specification of human-like semantic categories specifically tuned for human-like knowledge-based decision-making. Although ontology-based user profile model is being developed from recent times, to present days, it has become the mature approach to user profile modeling. The paper presents a critical survey on current state-of-the-art in ontology-based user profiling for 3G RecSys. In this survey, the special attention is paid to specific properties of source information to be represented in ontology model of user profile and corresponding technologies proposed.

Key words: *recommender systems, ontologies, user profile, personalization, context-aware recommendations, folksonomies, network context, automation of ontology development.*

References

[1] ***Tuzhilin, A.*** Keynote presentation at International Conference on Data Mining (ICDM 2013) / A. Tuzhilin. - Dallas, Texas, December, 2012.

[2] ***Cao, L.*** Actionable knowledge discovery and delivery / L. Cao // In: WIREs Data Mining and Knowledge Discovery, Volume 2, March/April 2012, John Willey & Sons , Inc., 2012. – P. 149–163.

[3] ***Argyri, C.*** Actionable Knowledge / C. Argyri // The Oxford Handbook of Organization Theory. Edited by Knudsen C. and Tsoukas H. - Oxford University Press, 2005.

[4] Recommender Systems Handbook / Ricci F., Rokach L. and Shapira B (Eds.). - Springer, 2011. – 842 P.

[5] ***Jannach, D.*** Tutorial: Recommender Systems / D. Jannach, G. Friedrich // International Joint Conference on Artificial Intelligence.(Beijing, August 4, 2013). Available at http://ijcai-11.iiia.csic.es/files/proceedings/ Tutorial%20IJCAI%202011%20Gesamt.pdf

[6] ***Segaran, T.*** Programming Collective Intelligence / T. Segaran // O'Reilly, 2006.

[7] ***Adomavicius, G.*** Incorporating contextual information in recommender systems using a multidimensional approach / G. Adomavicius, R. Sankaranarayanan, S. Sen, A. Tuzhilin // ACM Transactions on Information Systems (TOIS). – 2005 - 23(1). - P. 103–145.

[8] ***Adomavicius, G.*** Context-Aware Recommender Systems / G. Adomavicius, B. Mobasher, F. Ricci, A. Tuzhilin // AI Magazine, FALL 2011. - P. 67–80.

[9] ***Adomavicius, G.*** Context-Aware Recommender Systems / G. Adomavicius, A. Tuzhilin // In Ricci F., Rokach L, Shapira B,·Kantor P (Eds.). Recommender Systems Handbook, Springer, 2011. - P. 217–256.

[10] ***Gauch, S.*** Ontology-Based Personalized Search and Browsing / S. Gauch, J. Chaffee, A. Pretschner // ACM Web Intelligence and Agent System. - 2003. - Vol. 1. - No. 3/4. - P. 219–234.

[11] ***Leung, K.W.T.*** Deriving Concept-Based User Profiles from Search Engine Logs / K.W.T. Leung, D.L. Lee // IEEE Transaction on Data and Knowledge Engineering. - 2010. - Vol. 22. - No. 7. - P. 969–982.

[12] ***Liu, F.*** Personalized Web Search by Mapping User Queries to Categories / F. Liu, C. Yu, W. Meng // In Proc. of Intern. Conf. on Information and Knowledge Management (CIKM), 2002.

[13] ***Xu, Y.*** Privacy-Enhancing Personalized Web Search / Y. Xu, K. Wang, B. Zhang, Z. Chen // Proceedings of World Wide Web (WWW) Conference, 2007. - P. 591–600.

[14] Open Directory Project, http://www.dmoz.org/

[15] ***Costa, A.C.*** Cores: Context-aware, ontology-based recommender system for service recommendation / A.C. Costa, R.S.S. Guizzardi, J.G.P. Filho // In Proc. 19-th Intern. Conf. on Advanced Information Systems Engineering (CAISE07). 2007.

[16] ***Middleton, S.E.*** Ontological user profiling in recommender systems / S.E. Middleton, N.R. Shadbolt, D.C. de Roure // ACM Transaction on Information Systems. - 2004. - 22(1). - P. 54-88.

[17] ***Trajkova, J.*** Improving Ontology-Based User Profile / J. Trajkova, S. Gauch // RIAO, 2004. - P. 380-390.

[18] ***O'Hara, K.*** The AKT Manifesto / K. O'Hara, N. Shadbolt, S. Buckingham. - http://citeseer.uark.edu:8080/citeseerx/ showciting?cid=3895624

[19] ***Bizer, C.*** Linked Data – The Story So Far / C. Bizer, T. Heath, T. Berners-Lee // International Journal on Semantic Web and Information Systems (IJSWIS), 2009.

[20] ***Peña, P.*** Collective Knowledge Ontology User Profiling for Twitter / P. Peña, R. del Hoyo, J. Vea-Murguía , C. González, S. Mayo // 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.

[21] OpenDNS cloud websites tagging: http://community.opendns.com/domaintagging/

[22] DBpedia–a crowd-sourced community effort to extract structured information from Wikipedia: http://dbpedia.org

[23] The Apache Cassandra database: http://cassandra.apache.org/

[24] Virtuoso, a grade multi-model data server: http://virtuoso.openlinksw.com/

[25] NoSQL data bases: http://ru.wikipedia.org/wiki/NoSQL

[26] Pellet reasoning server: http://clarkparsia.com/pellet/

[27] ***Zeng, Y.*** User Interests: Definition, Vocabulary, and Utilization in Unifying Search and Reasoning / Y. Zeng, Y. Wang, Z.S. Huang, D. Damljanovic, Zh. Ning, C. Wang // In An A. et al. (Eds.): Active Media Technology 2010, Lecture Notes in Computer Science, vol. 6335, Springer, 2010. - P. 98–107.

[28] ***Su, Z.G.*** Research on Personalized Recommendation Algorithm Based on Ontological User Interest Model / Z.G. Su, J. Yan, H.F. Ling, H.S. Chen // J. of Computational Information Systems. – 2012. - Vol. 8. - No 1. - P. 169–181. Available also at http://www.Jofcis.com/

[29] ***Cantador, I.*** Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations / I. Cantador, M. Szomszor, H. Alani, M. Fernández , P. Castells // In Proc of 1st Intern. Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008), Tenerife, Spain. 2008.
[30] Delicious: Social Bookmark manager: http://ru.wikipedia.org/wiki/Delicious
[31] WordNet: http://ru.wikipedia.org/wiki/WordNet
[32] ***Miller, G.A.*** WordNet: A Lexical Database for English / G.A. Miller // Communications of the Association for Computing Machinery. – 1995. - 38(11). - P. 39-41.
[33] ***Castells, P.*** An Adaptation of the Vector-Space Model for Ontology-based Information Retrieval / P. Castells, M. Fernández, D. Valle // IEEE Transactions on Knowledge and Data Engineering. – 2007. - 19 (2). - P. 261-272.
[34] ***Vallet, D.*** Personalized Content Retrieval in Context Using Ontological Knowledge / D. Vallet, P. Castells, M. Fernández, P. Mylonas, Y. Avrithis //IEEE Trans. on Circuits and Systems for Video Technology. – 2007. - 17(3). - P. 336-346.
[35] ***Alfonseca, E.*** The Wraetlic NLP Suite / E. Alfonseca, A. Moreno-Sandoval, J.M. Guirao, M. Ruiz-Casado // In Proc. of the 5th Intern. Conf. on Language Resources and Evaluation. 2006. - P. 2277–2280. Available at http://www.lrec-conf.org/proceedings/lrec2006/
[36] Lucene: An Open Source Information Retrieval Library: http://lucene.apache.org/
[37] ***García-Silva, A.*** Social Tags and Linked Data for Ontology Development: A Case Study in the Financial Domain / A. García-Silva, L.J. García-Castro, A. García // In Proceedings of 4–th International Conference on Web Intelligence, Mining and Semantics (June 2–4, 2014, Thessaloniki, Greece).
[38] ***Jaschke, R.*** Discovering shared conceptualizations in folksonomies / R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, G. Stumme// Web Semantics Science Services and Agents on the World Wide Web. - 2008. - 6(1). - P. 38-53.
[39] ***Garcia-Silva, A.*** Review of the state of the art: discovering and associating semantics to tags in folksonomies / A. Garcia-Silva, O. Corcho, H. Alani, A. Gomez-Perez // The Knowledge Engineering Review. - 2012. - 27(2). - P 57-85.
[40] ***Kadima, H.*** Toward ontology-based personalization of a Recommender System in social network / H. Kadima, M. Malek // International Journal of Computer Information Systems and Industrial Management Applications. - 2013. - Vol. 5. - P. 499-508.
[41] ***Ma, Y.F.*** User Interests Modeling Based on Multi-source Personal Information Fusion and Semantic Reasoning / Y.F. Ma, Y. Zeng, R. Xu, Zh Ning // In Zhong N., Callaghan V., Ghorbani A., Hu B. (Eds.): Active Media Technology-2011, Lecture Notes in Computer Science, vol. 6890, Springer, 2011. - P. 195–205.
[42] ***Varshney, P.K.*** Multisensor data fusion / P.K. Varshney // Electronics & Communication Engineering J. – 1997. - 9(6). - P. 245–253.
[43] ***Ganter, B.*** Formal concept analysis: foundations and applications / B. Ganter, R. Wille. - Springer, 1999.
[44] ***Gorodetsky, V.*** Agent-based Customer Profile Learning in 3G Recommending Systems / V. Gorodetsky, V. Samoylov, O. Tushkanova // In: Proc. of the 9-th Intern. Workshop “Agent and Data Mining Interaction” (ADMI - 2014) associated with International Conference “Autonomous Agents and Multi-agent Systems” (AAMAS - 2014) (Paris, May 5–9 2014). To be also published in Post Proceedings of the Workshop as a volume of Lecture Notes in Computer Science, Springer, 2014.

Сведения об авторах

Городецкий Владимир Иванович, 1937 г. рождения. Окончил Ленинградскую военно-воздушную инженерную академию им. А.Ф. Можайского (1960) и математико-механический факультет Ленинградского госуниверситета (1970), д.т.н. (1973), профессор (1991). Гл. научный сотрудник лаборатории интеллектуальных систем Санкт-Петербургского института информатики и автоматизации РАН. Член Российской ассоциации искусственного интеллекта, IEEE Computer Society, International Society of Information Fusion (ISIF), International Foundation for Autonomous Agents and Multi-agent Systems (IFAAMAS). Автор более 300 публикаций и нескольких монографий. Область научных интересов: искусственный интеллект, в частности, технология многоагентных систем и инструментальные средства, прикладные многоагентные системы, распределенное обучение, извлечение знаний из баз данных, анализ и объединение данных различных источников, P2P сети принятия решений и P2P методы извлечения знаний из данных, обработка больших данных, планирование и составление расписаний, алгоритмы улучшения изображений, получаемых с помощью мобильных устройств, рекомендующие системы. (http://space.iias.spb.su/ai/gorodetsky)

Vladimir Ivanovich Gorodetskiy, Professor of Computer Science, Chief Scientist of Intelligent Systems Lab. of St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, received MS degree in mechanics from the Military Air Force Engineering Academy in St. Petersburg (1960) and MS degree in mathematics from Mathematical and Mechanical Department of the St. Petersburg State University (1970); Ph.D and and Doctor of Technical Sciences (1973) Science. Published more than 300 journal and conference papers and several books. Current scientific interests: Intelligent Data Analysis, Information Fusion, P2P Data Mining and Machine Learning, Multi-Agent Systems Technology and Software Tools, Agent-based Applications (Air Traffic Control, Intelligent Logistics, etc.), Recommender systems, Mobile Image Enhancement. Web site: http://space.iias.spb.su/ai/gorodetsky

Тушканова Ольга Николаевна, 1988 г. рождения. Получила степень магистра техники и технологии в области «Системный анализ и управление» в Южном федеральном университете, г. Ростов-на-Дону (2011). Аспирант, мл. научный сотрудник Санкт-Петербургского института информатики и автоматизации РАН. Область научных интересов: интеллектуальный анализ данных и извлечение знаний, многоагентные системы, рекомендующие системы, облачные технологии, онтологии.

Olga Nikolaevna Tushkanova (b. 1988) received MS degree in engineering and technology from Southern Federal University, Rostov-on-Don (2011). Currently is a graduate student and a junior researcher at the St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences. Current scientific interests: data mining, multi-agent systems, recommender systems, cloud computing, ontologies, knowledge extraction technologies.