Н.В. КАРПОВ,
Е.В. ШАДРИНА

Разработка сервиса поиска экспертов для актуальных информационных событий

УДК 004.51:004.82:81'322

ФГАОУ ВПО
«Национальный
исследовательский
университет
«Высшая
школа экономики»,
г. Нижний Новгород

Аннотация. В данной работе предлагается новый способ разработки сервиса обмена знаниями в университетском кластере при помощи поиска компетентных экспертов. Способ основан на современном подходе к поиску экспертов при помощи тематического моделирования. Сервис был реализован в виде системы поддержки принятия решений под названием EXPERTIZE.

Введение

Появление и успешный рост новых форм сотрудничества между организациями и университетами, известные как инновационные или университетские кластеры [1], стало важным феноменом в развитых экономиках. Непрерывный обмен профессиональными знаниями между заинтересованными участниками инновационных кластеров играет важную роль в экономике знаний. Для этой цели университет без сомнения должен играть роль катализатора, который предоставляет экспертные знания. Основные проблемы и стратегические решения должны комментироваться, обсуждаться и быть в центре внимания многочисленных заинтересованных лиц, включая средства массовой информации и общество.

До настоящего момента больших успехов в вопросе тесной интеграции университетского сообщества и появляющиеся инновационных кластеров не достигнуто. Информационные связи не образуются, так как большинство мероприятий проводится внутри регулярных университетских структур, таких как инкубаторы или бизнес парки. Общение с бизнес экспертами и средствами

массовой информации показывает, что в современной турбулентной информационной среде именно парадигма обмена информацией и знаниями должна быть модернизирована. Модернизированный обмен информацией и знаниями должен помочь университетскому сообществу ответить на важные экономические или социальные явления, возникающие в открытой среде инновационной экономики знаний.

Традиционному университетскому сообществу в России не хватает быстрых и всесторонних методов анализа текущей информации о важных обсуждаемых темах и ключевых проблемах. В современной университетской практике главным образом используется ручной анализ средств массовой информации и интернет ресурсов. Дальнейшее распространение информации об интересующих событиях происходит через неэффективную иерархическую организационную структуру: от руководителей структурных подразделений к сотрудникам.

Мы считаем, что усовершенствованные методы автоматизированного управления знаниями составляют важнейшую научную основу для модернизации обмена знаниями. Специально разработанное сочетание автоматической обработки текста и онтологического представления знаний может улучшить качество анализа информации.

В нашем исследовании мы ограничили проблему обмена знаниями до задачи поиска компетентного эксперта в реальном времени. Эксперт сопоставляется с актуальным информационным событием, происходящим в открытой экономической среде. Мы решаем эту задачу с помощью нового метода, основанного на тематическом моделировании.

В отличие от программы Media-ILOG, которая использует семантическое сопоставление, предложенное Biling и др. [2], новизна нашего исследование состоит в применении современного эффективного подхода для анализа семантики на основе вероятностного тематического моделирования [3] к поиску русскоязычных экспертов. Алгоритмически это реализовано в системе поддержки принятия решений под названием EXPERTIZE. Программный комплекс EXPERTIZE имеет сервисно-ориентированную архитектуру. В настоящее время он

функционирует и регулярно обновляет данные о компетенциях экспертов, используя открытый интернет-портал НИУ ВШЭ. Интерфейс сервиса InfoPort [4] позволяет системе EXPERTIZE получать доступ в реальном времени к персональной информации сотрудников НИУ ВШЭ.

Особенностью подхода является то, что исходный текст можно сопоставлять не только с экспертом из нашей базы, но и с категорией научных интересов, которая является элементом онтологии. Экспериментальное исследование показывает, что система EXPERTIZE позволяет корректно сопоставлять актуальное событие с компетентным экспертом.

Обзор релевантных методов поиска экспертов

Наша задача - сопоставить экспертные знания с простым текстом новостей – это переплетается с обычной задачей поиска экспертов. В последнее время возник колоссальный интерес к вопросу поиска экспертов. С 2005 года конференция по информационному поиску (TREC 2005), секция Enterprise Track предоставляет исследователям платформу для экспериментального тестирования методов и технологий поиска экспертов [5]. Тестовая коллекция TREC Enterprise основана на текстах из открытых баз данных научных организаций, таких как W3C (World Wide Web Consortium) и CSIRO (Commonwealth Scientific and Industrial Research Organisation).

В книге Balog и др. [5] авторы освещают современные модели и алгоритмы в данной области. Они классифицируют подходы к поиску экспертов таким образом:

- модель на основе профиля;
- документальная модель;
- гибридные модели.

Модель поиска эксперта на основе его профиля предложена в статье Balog и De Rijke [6]. Навыки кандидата представлены в виде баллов по отношению к документам, которые относятся к заданной области знаний. Релевантность документа оценивается, используя стандартные техники языкового моделирования.

В других подходах, например, документальном методе поиска эксперта, не создается профиль каждого эксперта. Этот метод

использует отдельные документы для сопоставления кандидатов с запросами. Идея состоит в том, чтобы сначала найти документы, относящиеся к теме, а потом определить экспертов, ассоциированных с этими документами. Позже, Fang и Zhai [7] представили общую вероятностную модель поиска эксперта и показали, как документальная модель может быть адаптирована к этой схеме.

Интересный подход для поиска эксперта предложил В.А. Фомичев [8]. Подход позволяет выстроить и сравнить семантические представления профиля эксперта, используя К-репрезентацию и модель лингвистической базы данных.

Balog и др. [9] в своей работе для поиска экспертов использовали моделирование тематик. Вместо моделирования профиля кандидата или документа они строили модель для каждого входящего запроса и использовали эту модель для вычисления вероятности релевантности кандидата данному запросу. Их подход близок к документальному подходу, который часто используется в информационном поиске и основан на языковой модели. Судя по полученным ими результатам, эта модель работает хуже, чем модели на базе профиля и документальные модели. Основной причиной этому является разреженность моделей, построенных на основе запросов. Их определение тематик, при этом, отличается от нашего понимания. Термин «тема» в их работе относится к словам запроса, которые пользователи используют для поиска эксперта, тогда как в настоящей работе мы используем термин «тема» как набор понятий, который можно извлечь из совокупности слов в документе, используя алгоритм тематического моделирования. Существует много методов тематического моделирования документа, такие как Латентно-семантический Анализ (LSA) [10], Латентное размещение Дирихле (LDA) [3] и др.

Подход к тематическому моделированию основан на предположении мешка слов, то есть слова в документе независимы друг от друга и от порядка в тексте. Ведем обозначения:

D – множество документов;

W – множество уникальных слов в множестве D;

Z – множество скрытых тем.

Документы в множестве D независимые и неупорядоченные. Каждая скрытая тема имеет свое собственное распределение слов *P*(W), и каждый документ имеет распределение по скрытым темам *P*(Z).

В результате применения подхода мы получаем следующее распределение условных вероятностей:

- $P(w_i / z_k);\ i \in \overline{1,|\mathbf{W}|}, k \in \overline{1,|\mathbf{Z}|}$ обозначим как $P(W/Z)$.
- $P(z_k / d_n);\ k \in \overline{1,|\mathbf{Z}|}, n \in \overline{1,|\mathbf{D}|}$ обозначим как $P(Z/D)$.

Интересный подход на основе тематик предложен Момтаци и Науманн [11]. Этот подход показывает более высокую точность, чем известные подходы на основе профилей экспертов и документационного анализа. Документы не используются для осуществления поиска. Вместо этого мы используем эти документы только для обучения модели LDA. Потом, непосредственно в поиске, мы используем распределение $P(W/Z)$ для оценки распределения вероятностей $P(E/Z)$, так как E – это подмножество экспертов множества W.

В работе (17) исследователи демонстрируют, как использовать модель на основе тем с научной онтологией, где каждый документ имеет метку категории в научном классификаторе. C – множество категорий в научном классификаторе. Блай и Хаффман представляют каждую категорию c_j как распределение условных вероятностей $P(z_k / c_j);\ k \in \overline{1,|Z|}, j \in \overline{1,|C|}$ с помощью формулы:

$$P(z_k / c_j) = \frac{P(c_j / z_k)P(z_k)}{\sum_k P(c_j, z_k)} = P(Z / c_j) \qquad (1)$$

где, $k = \overline{1,|Z|}$, $P(c_j / z_k)$ и $P(z_k)$ могут быть получены из модели LDA. Так как $\sum_k P(c_j, z_k)$ является константой для разных c_j, а $P(z_k)$ это равномерное распределение, то имеем линейную зависимость

$$P(\mathrm{Z}/\mathrm{c}_\mathrm{j}) \propto P(\mathrm{c}_\mathrm{j}/\mathrm{Z}) \qquad (2)$$

Основываясь на проведенном исследовании работ разных авторов, считаем, что лучшим решением нашей задачи является использование модели на основе тематического моделирования [12]. Таким образом, мы должны проанализировать документы на русском языке, находящиеся в базе данных нашего университета. С

помощью подхода, описанного в работе [13], метод на основе тематического моделирования может быть применен для работы с научной онтологией. Таким образом, мы применяем новый эффективный алгоритм и существующие научные онтологии.

Методика

В предыдущей разработанной нами программе InfoPort [4] для решения проблемы поиска эксперта мы предложили переводить запрос пользователя в соответствующий запрос SPARQL (Protocol and RDF Query Language), который соотносился с набором данных в формате RDF (Resource Description Framework). Результатом запроса являлась соответствующая категория научного классификатора и ключевые слова. Алгоритм поиска системы InfoPort находил всех экспертов, которые соответствуют научному классификатору или ключевому слову.

В настоящей работе наша новая система EXPERTIZE работает автоматически: она воспринимает событие, выраженное в новости, как запрос и сопоставляет его с наиболее подходящим экспертом, который может высказать свое мнение по этой теме. Другими словами, мы ранжируем экспертов по степени их соответствия конкретному событию.

С одной стороны, мы имеем возможность извлечь семантическую информацию из текста, так как новости представлены в обычном текстовом формате. С другой стороны, каждый эксперт имеет опубликованные статьи или записи устных интервью и семинаров. Этот материал содержит богатую информацию об интересах и компетенциях персоны.

Существует несколько формальных моделей, подходящих для проведения контекстуального анализа, такие как: дистрибутивная модель [14] [15], латентный семантический анализ [16] [10] и латентное размещение Дирихле (LDA) [3]. В нашем исследовании мы используем последнюю модель.

В настоящее время есть несколько методов построения моделей LDA, иначе говоря, методов поиска параметров всех функций распределения в модели. Все методы аналогичны по структуре EM (Expectation-maximization) алгоритму. К ним относятся:

Байесовские вариационные методы [17];

семплирование Гиббса [18];

Expectation Propagation [19];

Из этих трех алгоритмов мы используем Байесовский вариационный метод как наиболее точный [17]. Он реализован в пакете Gensim.

Первым шагом нашего метода поиска эксперта является обучение модели на коллекции текстов. Мы вычисляем оценку двух дискретных функций распределения $P(w_i / z_k);\ i \in \overline{1,|\mathrm{W}|}, k \in \overline{1,|\mathrm{Z}|}$, $P(z_k / d_n);\ k \in \overline{1,|\mathrm{Z}|}, n \in \overline{1,|\mathrm{D}|}$.

На вход системы EXPERTIZE поступает новость, обозначим ее как d_0. Семантическое представление новости d_0 может быть вычислено с использованием построенной модели LDA, а именно: $P(z_k / d_0);\ k \in \overline{1,|\mathrm{K}|}$.

На втором шаге скрытые тематики Z используются для вычисления релевантности документа d_0, кандидатам из множества E и категориям из множества C. Оба множества E и C представлены как подмножества слов в модели LDA. Таким образом, $P(d_0 / \mathrm{E})$ и $P(d_0/\mathrm{C})$ вычисляются на основании скрытых тем, которые распределены по экспертам (E) и по научным категориям (C).

$$P(d_0 / \mathrm{E}) = \sum_{z \in \mathbf{Z}} P(d_0 / z, \mathrm{E}) P(z / \mathrm{E}) \tag{3}$$

$$P(d_0 / \mathrm{C}) = \sum_{z \in \mathbf{Z}} P(d_0 / z, \mathrm{C}) P(z / \mathrm{C}) \tag{4}$$

Принимая предположение о независимости между d_0 и E, C и предполагая что документ d_0 равновероятен с другими документами, имеем:

$$P(d_0 / \mathrm{Z}, \mathrm{E}) = P(d_0 / \mathrm{Z}, \mathrm{C}) = P(d_0 / \mathrm{Z}) = \\ = \frac{P(\mathrm{Z} / d_0) P(d_0)}{P(\mathrm{Z})} \propto P(\mathrm{Z} / d_0) \tag{5}$$

Используя (2) и (5) из (3) и (4) получаем следующие простые формулы:

$$P(d_0 / \mathrm{E}) \propto \sum_{z \in \mathbf{Z}} P(z / d_0) P(\mathrm{E} / z) \tag{6}$$

$$P(d_0 / \mathrm{C}) \propto \sum_{z \in \mathbf{Z}} P(z / d_0) P(\mathrm{C} / z) \tag{7}$$

Мы ранжируем экспертов E из списка сотрудников компании исходя из критерия максимума условной вероятности:

$$e_{\max} = \arg\max_{i \in 1,|E|} \left(P(d_0 / e_i) \right) \quad (8)$$

Для ранжирования категорий **C** из научного классификатора мы используем аналогичный подход. Далее выбираются наиболее вероятные категории и ассоциируются с экспертом.

Структура программы EXPERTIZE

Метод сопоставления актуального информационного события с экспертом из числа сотрудников университета был применен при разработке и внедрении системы EXPERTIZE. На верхнем уровне программы можно выделить следующие сервисы (рис. 2):

- Сбор данных;
- Моделирование данных;
- Хранилище данных;
- Графический интерфейс пользователя;
- Сопоставление данных.

Программа EXPERTIZE активно использует наш сервис InfoPort [12]. Этот семантический сервис предоставляет фактическую информацию о более чем трехстах сотрудниках Вышей Школы Экономики (НИУ ВШЭ – Нижний Новгород) в форме онтологии. Информация в InfoPort представлена в виде RDF триплетов. Триплеты включают иерархическую информацию, как в первоисточнике. Первый уровень - это список ученых в алфавитном порядке, второй – это ученые имеющие научные интересы и публикации, и третий – это тексты публикаций.

Компоненты программы EXPERTIZE могут быть классифицированы как онлайн и оффлайн сервисы. И те, и другие взаимодействуют с программой InfoPort через встроенный REST (Representational state transfer) интерфейс. Оффлайн сервисы работают регулярно для обновления информации. Онлайн сервисы работают по запросу, когда пользователь активирует их через веб-интерфейс.

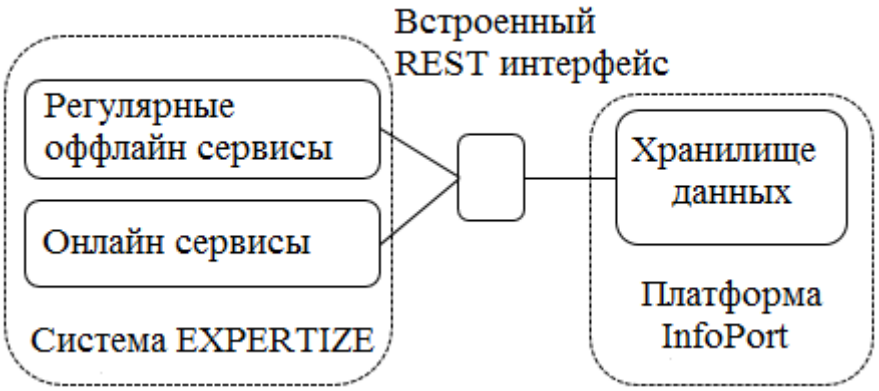

Рис. 1. Взаимодействие сервисов EXPERTIZE с платформой InfoPort.

Оффлайн обработка начинается с сервиса сбора информации диспетчером. Он делает запрос через REST-интерфейс в службу хранилища InfoPort, чтобы взять список ссылок URI (унифицированных идентификаторов ресурсов) научных работ. Так как каждая работа доступна онлайн, Сборщик находит ее по URI и извлекает характеристики работы, используя синтаксический анализатор XML (парсер). Характеристики научной работы включают: автора, название, аннотацию, ключевые слова, научные категории онтологии. Эта информация хранится в базе данных MySQL как временная исходная информация. Сервис Сбор данных использует язык программирования Python и библиотеку Lxml для обработки HTML страниц.

В сервисе Моделирование данных осуществляются следующая обработка данных:

- получение временной исходной информации;
- разбиение текста на слова;
- лемматизизация слов;
- индексирование слов с помощью словаря лемм;
- фильтрация слишком часто встречающихся в тексте слов (стоп-слова) или слишком редко встречающихся (использованные только один раз);
- индексирование авторов и научных категорий;
- формирование «мешков слов», используя леммы, авторов и категории;
- построение модели LDA с заданным количеством тем K.

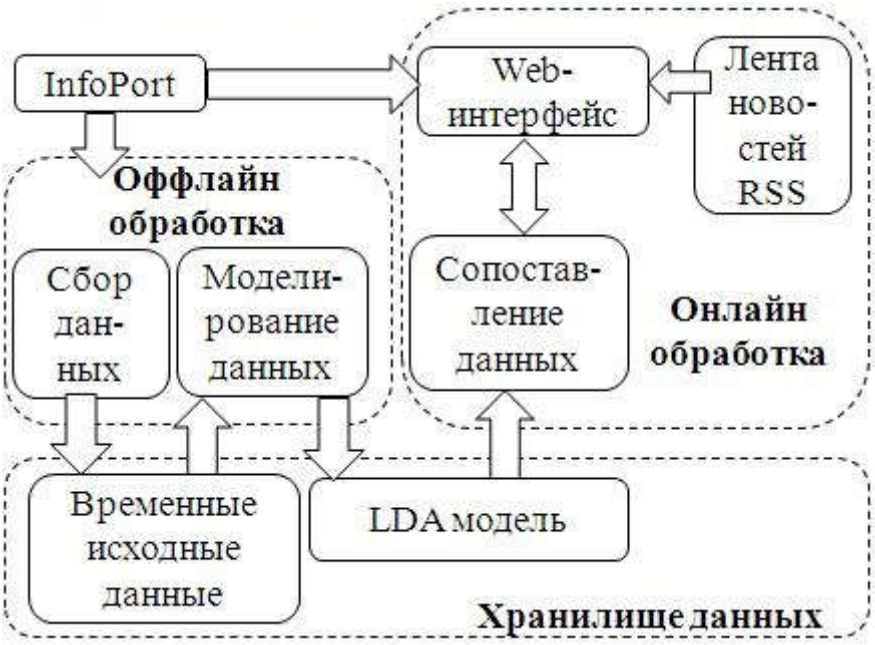

Рис. 2. Структура программы EXPERTIZE.

Онлайн обработка производится по запросу пользователя путем открытия графического веб-интерфейса. Веб-интерфейс активирует ленту новостей RSS, которая получает и показывает 10 последних новостей из ленты RSS и пустое текстовое поле. Пользователь может выбрать одну из 10 новостей или набрать вручную текст в поле. После того, как пользователь обозначил запрос, интерфейс передает его сервису Сопоставление данных. В свою очередь этот компонент системы выполняет семантический поиск онлайн. Семантическое представление новости сопоставляется с семантическим представлением научных категорий и экспертов с помощью формул (8) и (9) и выбираются первые 3 кандидатуры. Таким образом, сервис Сопоставление данных возвращает графическому интерфейсу три ссылки на личные страницы сотрудников университета.

Для обеспечения удобного для пользователя представления полученных данных графический интерфейс делает запрос в сервис InfoPort и оттуда получает дополнительные данные по выбранным ссылкам: полное имя, URL-ссылку на фото эксперта, департамент, где работает эксперт.

Экспериментальная часть

Оценка предложенного нами метода и системы EXPERTIZE была проведена эмпирическим путем на основе выборки из 100 текстов. Для этого мы выбираем эксперта из нашей университетской базы. В этой базе содержатся более трехсот профессоров и исследователей из НИУ ВШЭ – Нижний Новгород. В соответствии с областью научных интересов выбранного сотрудника мы находим новость, которую эксперт способен прокомментировать и вносим их в программу EXPERTIZE. Если этот эксперт появляется в листе кандидатов, предложенной системой, мы считаем попытку поиска успешной.

Рассмотрим пример. Выбираем в качестве эксперта Дмитрия Сидорова, его научные интересы включают ряд тем, среди которых:

w_1 – инновационные проекты

w_2 – венчурные инвестиции

w_3 – оценка инновационного потенциала

и др.

Каждая тема научной области кодируется как одно слово, и у нас есть заранее созданная таблица с распределением вероятностей слов W в латентных темах Z: $P(w_i / z_k);\ i \in \overline{1,|\mathrm{W}|}, k \in \overline{1,|\mathrm{Z}|}$. В ней обычно мало элементов со значением больше 0.

Таблица 1

Пример распределения вероятностей слова w_1 в скрытых темах

$$P(w_1 / z_k); k \in \overline{1,|\mathrm{Z}|}$$

	z_1	z_2	…	z_{58}	..	z_{200}
w_1	0	0.04		0.1		0

Мы находим новость под названием «Яндекс платит за BIG Data», которую выбранный эксперт способен прокомментировать. Это статья об инвестициях российского IT-гиганта в израильскую молодую компанию. Статья поступает на вход сервису Сопоставление данных, где она конвертируется в распределение вероятности по скрытым темам [2], используя заранее построенную модель LDA. Количество скрытых тем мы установили равным 200.

Таблица 2

Пример распределения вероятностей тем Z в тексте d_0 –

$$P(z_k / d_0);\ k \in \overline{1,|\mathrm{K}|}$$

	z_1	z_2	…	z_{58}	.	z_{200}
d_0	0	0.21		0.058		0.034

Далее, используя формулу (10), алгоритм перебирает всех экспертов из существующего набора **E** и находит вероятность $P(d_0 / \mathrm{E})$. В результате система выдает трех экспертов с максимальной величиной. Результат представлен на рис. 3.

Рис. 3. Графический интерфейс пользователя системы EXPERTIZE.

Так как наша цель, эксперт Сидоров Дмитрий, представлен в выходных данных, считаем эту попытку успешной.

Для сравнения наших результатов с аналогичными системами, мы используем в качестве метрики качества среднюю точность работы системы – MAP (Mean average precision).

$$MAP = \frac{1}{K}\sum_{k=1}^{K} P(k);$$
$$P(k) = \frac{\left|R \cap \hat{R}_k\right|}{\left|\hat{R}_k\right|} \quad (11)$$

где R – все множество релевантных экспертов в базе данных, $\hat{R}_k$ – набор из k экспертов, выбранных системой как релевантные. В среднем 43 попытки оказались успешными из 100 проделанных опытов, то есть MAP равна 0,43.

Таблица 3.

Экспериментальные результаты с разными моделями онтологических сопоставлений

Показатель качества работы системы	**Значение**
MAP, коллекция НИУ ВШЭ	0,43
MAP, English TREC 2006	0,471
MAP, English TREC 2005	0,248

Наша реализация системы поиска русскоязычных экспертов позволяет получать результат сравнимый по точности с результатами из работы [11]. При этом, архитектура системы позволяет использовать ее для облегчения обмена знаниями. Средняя точность MAP в диапазоне 0,4-0,5 считается приемлемой в данной задаче.

Заключение

В данной статье мы представили новый подход для разработки сервиса быстрого обмена знаниями в инновационных кластерах, основанный на поиске экспертов. Для доступа к компетенциям потенциальных экспертов предложенный метод использует открытые Интернет-ресурсы и существующие онтологические сервисы, такие как InfoPort [4].

В процессе нашего исследования мы проанализировали качество работы нового метода поиска экспертов, основанного на тематическом моделировании, для русского языка. Результат работы включает в себя программное решение для сопоставления соответствующего университетского эксперта и актуальных информационных событий, происходящих в открытой среде. Это

решение позволяет сопоставлять в реальном времени новости, размещенные в сети Интернет, и сферы интересов сотрудников университета, с последующим быстрым уведомлением о возможном участии подходящих сотрудников в интервью, информационных программах и дискуссиях.

Программа поддержки принятия решений под названием EXPERTIZE была разработана для практического применения предложенного метода. Первый опыт использования программы EXPERTIZE показывает, что она справляется с поставленными задачами. Используя модель на основе тематик, предложенную Momtazi и Naumann [11], мы получили среднюю точность MAP равную 0,43. Тот же подход на англоязычных коллекциях TREC 2005 и TREC 2006 показал среднюю точность MAP равную 0,248 и 0,471 соответственно. Таким образом точность программы EXPERTIZE немногим ниже, чем точность, полученная по TREC 2006. Оценка других показателей качества системы EXPERTIZE, например, полнота и F-мера (F-measure), для нашего исследования не так интересна, потому, что рядовому пользователю не нужен полный набор разнообразных экспертов. Один или два подходящих эксперта вполне достаточно для облегчения обмена знаниями.

Предложенный подход позволяет эффективно сопоставлять информационное событие не только с самими экспертами, но и с научными категориями. Это позволяет нам, при помощи многоязычной научной онтологии, осуществить поиск англоязычных экспертов, имея запрос на русском языке или наоборот. Такое исследование мы планируем осуществить в нашей следующей научной работе.

Литература

1. Asheim B., Cooke P., Martin R.; Clusters and regional development: Critical reflections and explorations, Econ. Geogr. 2008, Т. 84, № 1. С. 109–112.

2. Billig A., Blomqvist E., Lin F.; Semantic matching based on enterprise ontologies, On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS. Springer, 2007, С. 1161–1168.

3. Blei D.M., Ng A.Y., Jordan M.I.; Latent dirichlet allocation, J. Mach. Learn. Res. 2003. Т. 3. С. 993–1022.

4. Babkin E., Karpov N., Kozyrev O.; Towards Creating an Evolvable Semantic Platform for Formation of Research Teams, Perspectives in Business Informatics Research. Springer, 2013, С. 200–213.

5. Balog K. и др.; Expertise Retrieval, Foundations and Trends in Information Retrieval. 2012, Т. 6, № 2-3. С. 127–256.
6. Balog K., De Rijke M.; Determining Expert Profiles (With an Application to Expert Finding)., IJCAI. 2007, Т. 7. С. 2657–2662.
7. Fang H., Zhai C.; Probabilistic models for expert finding, Advances in Information Retrieval. Springer, 2007, С. 418–430.
8. Fomichov V.; Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Springer, 2009, Т. 27.
9. Balog K. и др.; Broad expertise retrieval in sparse data environments, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007, С. 551–558.
10. Dumais S.T.; Latent semantic analysis, Annu. Rev. Inf. Sci. Technol. 2004, Т. 38, № 1. С. 188–230.
11. Momtazi S., Naumann F.; Topic modeling for expert finding using latent Dirichlet allocation, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2013, Т. 3, № 5. С. 346–353.
12. Klimova N., Litvintseva M.; Universities Innovation Clusters: Approaches for National Competitiveness Paradigm, European Journal of Social Sciences. 2011, Т. 19, № 1. С. 160–162.
13. Zhu H. и др.; Towards expert finding by leveraging relevant categories in authority ranking, Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011, С. 2221–2224.
14. Baroni M., Lenci A.; Distributional memory: A general framework for corpus-based semantics, Computational Linguistics. 2010, Т. 36, № 4. С. 673–721.
15. Turney P.D.; Similarity of semantic relations, Computational Linguistics. 2006, Т. 32, № 3. С. 379–416.
16. Landauer T.K., Foltz P.W., Laham D.; An introduction to latent semantic analysis, Discourse Process. 1998, Т. 25, № 2-3. С. 259–284.
17. Blei D.M., Hoffman M.D.; Online Learning for Latent Dirichlet Allocation, Advances in Neural Information Processing Systems. 2010,
18. Griffiths T.L., Steyvers M.; Finding scientific topics, Proc. Natl. Acad. Sci. U. S. A. 2004, Т. 101, № Suppl 1. С. 5228–5235.
19. Minka T., Lafferty J.; Expectation-propagation for the generative aspect model, Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 2002, С. 352–359.