Модель предметной области на основе сервиса Google Scholar Citations

© Д. В. Ландэ
Институт проблем регистрации информации НАН Украины,
Киев, Украина
dwlande@gmail.com

Аннотация

Предлагается методика построения терминологических сетей – моделей предметных областей на основе зондирования информационных сетей. Как такая сеть рассматривается сеть понятий, соответствующих тегам сервиса Google Scholar Citations. Предложенный подход можно применять для многих областей науки.

1 Задача создания модели предметной области

В настоящее время задача создания моделей предметных областей все еще остается актуальной. Под моделью предметной области, в частности, понимают специальным образом сформированную сеть понятий (отраслевую онтологию). Построение большой отраслевой онтологии – сложная научно-практическая проблема [1], [3]. Первый этап этого процесса – построение терминологической основы онтологии и определение семантических связей [5].

В этой работе представляется подход к созданию модели предметной области (искусственный интеллект) на основе зондирования большой информационной сети. Как такая сеть рассматривается сеть понятий, которые отражаются в тегах наукометрического сервиса Google Scholar Citations (http://scholar.google.com/citations). На рис. 1 приведен фрагмент интерфейса страницы сервиса Google Scholar Citations, соответствующий заданному заранее тегу machine_learning (машинное обучение).

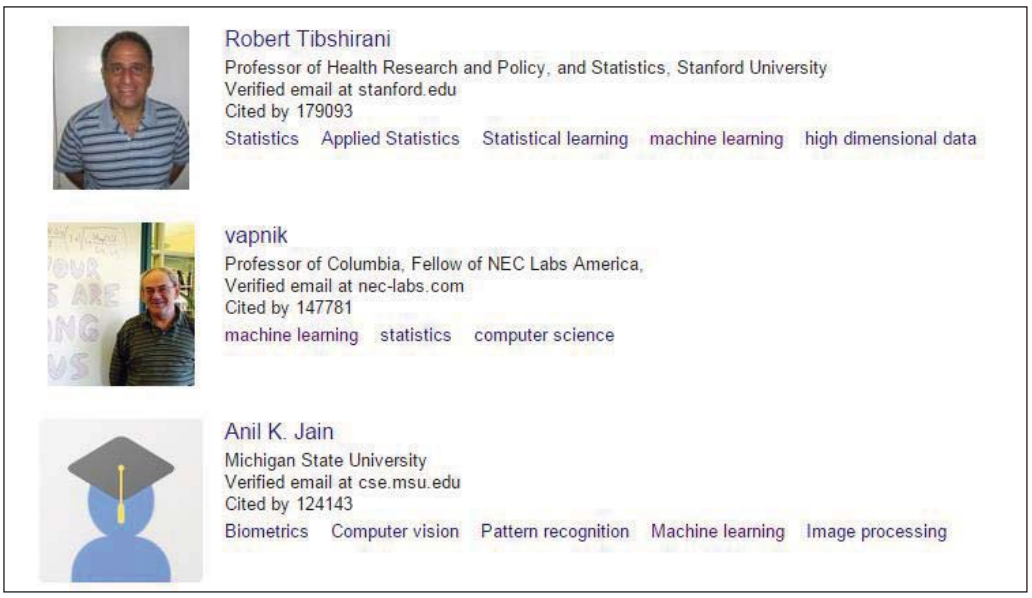


Рис. 1. Интерфейс страницы сервиса Google Scholar Citations

На интерфейсе, соответствующем данному тегу (label: machine_learning) постранично в ранжированном виде отображаются имена ученых, которые отметили свою деятельность этим тегом, а также другие теги, приписанные ими (например, Robert Tibshirani определил для себя еще Statistics, Applied Statistics, Statistical learning, machine learning, high dimensional data). Множество

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

тегов образуют сеть, производную от биграфа «ученый-теги». Эту сеть можно рассматривать как некоторую онтологическую модель предметной области. Узлы в этой сети соответствуют понятиям, маркированным тегами, а связи – некоторую семантическую связь между ними.

Конечно, теги указанные отдельными учеными могут относиться к различным отраслям науки, однако, предварительно проведенные исследования показывают, что на достаточно репрезентативной выборке (порядка сотни тегов), небольшая частота нетематических тегов обеспечивает их автоматическое «отсеивание» с помощью соответствующего алгоритма.

Целью работы является описание теоретических принципов и методологии автоматизированного формирования модели предметной области, в частности, области искусственного интеллекта в целом путем зондирования большой информационной сети. Для достижения этой цели применяется специальный алгоритм сканирования ресурсов сервиса Google Scholar Citations с целью получения репрезентативного набора тегов (обозначений понятий) как основы будущей онтологии. Под зондированием информационных сетей понимается выборка небольшого объема важнейшего содержания из больших информационных сетей, которые по технологическим причинам не подлежат полному сканированию.

2 Описание модели

При построении сетей тегов целесообразно применять модели, уже апробированные на пиринговых сетях (peer to peer, P2P – равный с равным), основанных на равноправии участников. В таких сетях отсутствуют выделенные серверы, а каждый узел (peer) является как клиентом, так и сервером. Во многих случаях P2P являются наложенными (оверлейными) сетями, которые используют существующие транспортные протоколы сети Интернет. Пиринговые сети состоят из узлов, каждый из которых взаимодействует лишь с некоторым подмножеством других узлов сети (ввиду ограниченности ресурсов).

Рассмотрим сеть, которую будем считать пиринговой, к тому же соединенной с глобальной компьютерной сетью, рассматриваемой в качестве внешней среды. Для поиска необходимых данных в таких сетях применяется несколько моделей. В модели "широкого первичного поиска" (Breadth First Search, BFS) запрос из некоторого стартового узла адресуется ко всем соседям (ближайшим по некоторым критериям). Когда некоторый другой узел получает запрос, выполняется поиск в его локальном индексе и в случае успеха возвращает результат. В противном случае запрос передается по сети далее. В результате успешного поиска формируется сообщение-отзыв (QueryHit), которое включает информацию о релевантных тегах, и доставляется по сети стартовому узлу. Другой алгоритм, так называемый "интеллектуальный поисковый механизм" (Intelligent Search Mechanism, ISM) обеспечивает улучшение скорости и эффективности поиска информации за счет минимизации количества сообщений между узлами и количества узлов, опрашиваемых для каждого поискового запроса [4], [6]. Чтобы достичь этого, для каждого запроса оцениваются лишь такие узлы, которые в наибольшей мере соответствуют запросу.

Именно модель, близкую к ISM будем рассматривать в этой работе.

Зондирование опорной модельной сети осуществляется по такому алгоритму:

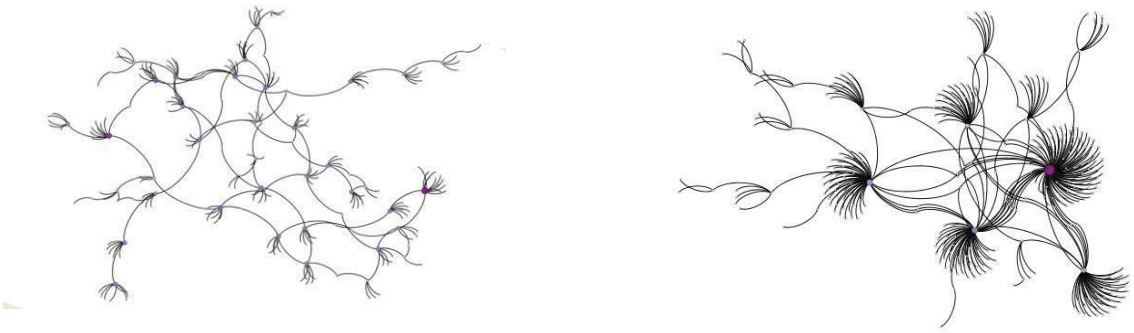
- Выбирается определенное количество узлов опорной (зондируемой) сети, определяемых как базовые для новой сети, соответствующей результатам зондирования.
- Для каждого из рассматриваемых узлов опорной сети определяются смежные с ним узлы ("соседи"), которые добавляются к создаваемой сети с результатами зондирования.
- От текущего узла опорной сети осуществляется переход к соседнему узлу, имеющему наибольшую степень.
- Если имеет место "зацикливание" (выбирается узел, к которому уже был осуществлен переход по этому алгоритму), происходит переход к следующему по степени соседнему узлу. Если таких узлов не осталось – осуществляется переход к пункту 2.
- Если перечень базовых узлов завершен, считается, что сеть, соответствующая результатам зондирования, построена

При моделировании приведенный алгоритм применялся для двух самых распространенных модельных сетей Erdős-Rényi (ER) и Barabási-Albert (рис. 2) [2]. Известно, что модель ER – это случайная сеть, которая строится следующим образом: множество из N изначально не соединенных узлов попарно объединяют с вероятностью p. В результате создается сеть приблизительно с $pN(N-1)/2$ случайно выбранными связями.

Модель BA – одна из нескольких моделей сетей со степенным распределением степеней узлов (так называемых, безмасштабных сетей). Эта модель учитывает как рост сети (динамику), так и принцип преимущественного присоединения, который заключается в том, что чем больше связей имеет узел, тем более вероятно для него создание новых связей со вновь образуемыми узлами. Узлы с большей степенью имеют большую вероятность присоединения (создания новых связей) к новым узлам.

Следует заметить, что безмаштабними являются наиболее популярные реальные сети, такие как веб-пространство с гиперссылками, социальные сети, сети слов в литературных произведениях, сети

протеинов, и т.п. [5] Автором изначально предполагалось, что сети понятий, естественным образом формируемые участниками сетевых сервисов тоже обладают свойством безмасштабности, что не всегда можно это проверить, не имея всеобъемлющей информации. Если сеть такая сложная и большая, как, например, Google Scholar Citations, на помощь может прийти зондирование, в результате которой выполняется построение некоторой новой сети, лишь частично совпадающей с исходной, большей по объему. Отметим, что результаты любого зондирования не всегда верно отображают природу большой исследуемой сети – результаты во многом зависят именно от алгоритма этой процедуры, вместе с тем, оно может служить базой для гипотез о структуре большой сети.



а *б*

Рис. 2. Пример сети, построенной зондированием модельных сетей:
а – Erdős-Rényi; *б* – Barabási-Albert

Исходя из информации о том, что все известные большие сети цитирования, соавторства и т.п. обладают свойством безмасштабности, т.е. в чем-то близки по структуре сети Barabási-Albert, автором изучались модели, одна из которых базировалась на алгоритме BA. От этой модели принципиально отличаются случайные сети Erdős-Rényi, которые также изучались для сравнения. Визуально качественные результаты зондирования сетей ER и BA с близкими параметрами (1000 узлов, около 2000 связей) приведены на рис. 2. Сравнение показывает, что связанные области (ветки), соответствующие отдельным понятиям в первом случае достаточно длинные, а узлов, по которым следует маршрут зондирования больше, чем во втором, более интересном для нас, случая. В рамках данного исследования более важны именно качественные результаты, вид связанных цепочек, которыми моделируются ветки понятий. Следует отметить, что реальным сетям присущий еще и феномен "клуба богатых" (Rich Clube), который обусловливает более плотную связанность наибольших узлов. Поэтому изначально предусматривалось, что приведенный алгоритм при зондировании реальной сети будет быстро «зацикливаться» (и, соответственно, прерываться), что приведет к еще большему сокращению веток понятий.

Именно на основании результатов качественного моделирования был сделан вывод о возможности формирования небольших связанных веток тегов, соответствующих понятиям, интересующим пользователей сервиса Google Scholar Citations.

3 Зондирование сети Google Scholar Citations

Приведенный выше алгоритм, которой применялся к модельным сетям, был адаптирован к реальной сети тегов сервиса Google Scholar Citations следующим образом:

- Экспертным путем определяется небольшой перечень базовых тегов (ключевых слов, соответствующих наиболее важным понятиям).
- Выбирается тег из определенного экспертами перечня.
- Открываются страницы веб-сервиса, соответствующие этому тегу (максимальное количество таких страниц параметрически ограничивается заранее).
- К создаваемой сети добавляются все теги, содержащиеся на выбранных страницах (соседние теги).
- Из соседних тегов выбирается тот, на страницы которого планируется перейти для дальнейшего анализа. Этот тег с наибольшей степенью среди соседних тегов, который также удовлетворяет тематике выбранной предметной области и не входит в состав тех тегов, к страницам которых уже был осуществлен переход.
- Если такой тег выбран, то происходит переход к пункту 3.

- Если такого тега не существует, но перечень базовых тегов не завершен, то осуществляется переход к следующему базовому тегу из начального перечня, т.е. переход к пункту 2. Иначе считается, что сеть зондирования построена.

В соответствии с приведенным алгоритмом процесс зондирования сети, начиная с определенного узла, прекращается при «зацикливании», т.е. когда в соответствии с алгоритмом происходил переход к уже пройденному тегу, а также при отклонении оставшихся соседних тегов от основной тематики (это определяется экспертами при автоматизированном зондировании или с учетом лексического состава тегов при полностью автоматическом сканировании). При этом само «зацикливание» является признаком перехода к следующему базовому тегу или завершению процесса зондирования.

Формирование базового стартового перечня узлов-понятий и правил отбора «конечных» узлов выполняется экспертами в предметной области.

Для построения модели предметной области (в рассматриваемом примере для области искусственного интеллекта) экспертным путем были определены базовые теги на английском языке: artifical_intelligence, neural_networks, machine_learning, heuristic_search, evolutionary_computation и др.

На рис. 3 приведен пример сети понятий предметной области, построенной в соответствии с приведенным алгоритмом по указанным базовым тегам. Конечно, состав узлов полученной сети существенно зависит от введенных экспертами начальных тегов. Именно потому данный поход нельзя считать полностью автоматическим.

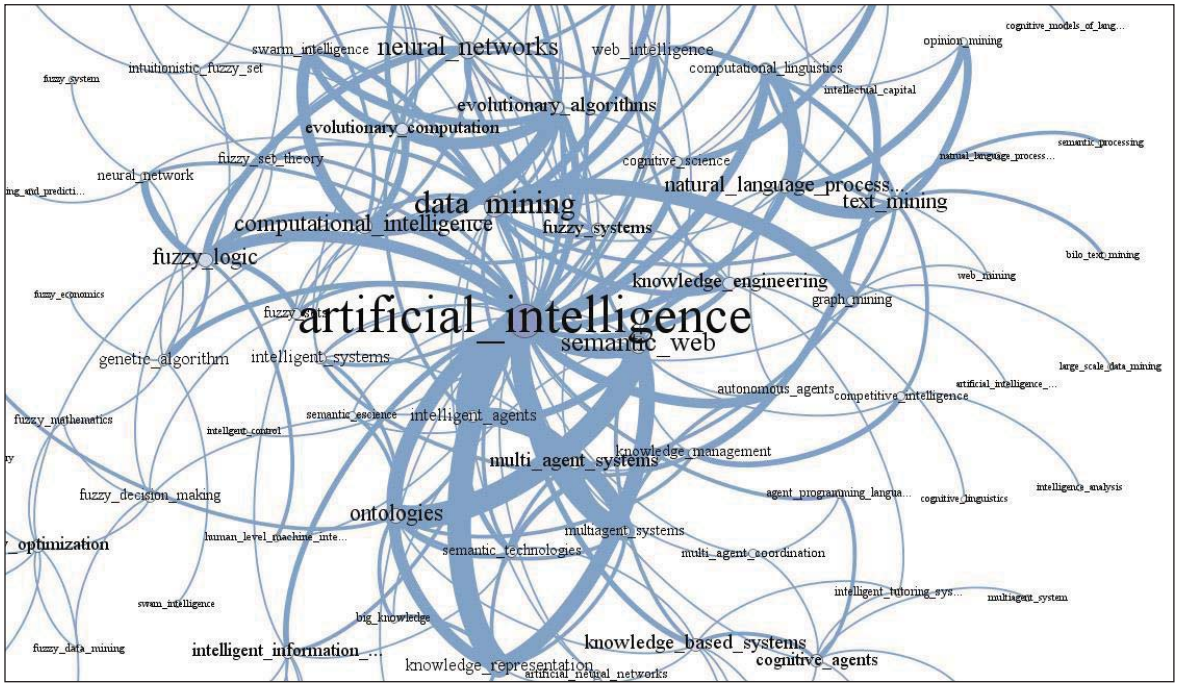


Рис. 3. Фрагмент сети понятий

Построенная сеть понятий оказалась связной. При количестве базовых тегов 20, общее количество узлов-тегов, которые были охвачены алгоритмом, составили 122, а количество нетерминальных узлов – лишь 65. Распределение степеней этих узлов, приведенное на рис. 4, свидетельствует об отсутствии степенного распределения, т.е. приведенный алгоритм зондирования, скорее всего не сохранил предполагаемого распределения степеней узлов базовой сети. Средняя длина ветви понятий составляет примерно 6.

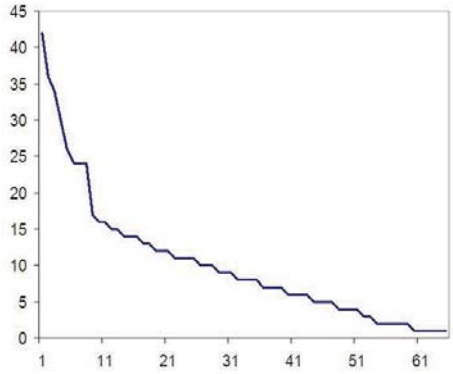


Рис. 4. Распределение степеней узлов-тегов сети понятий

Выводы

В предложенной модели предметной области как онтологические связи применяются связи между областями интересов отдельных ученых.

Фактически рассматривается компактификация биграфа «ученый – области науки, его интересующие».

Предложен и реализован подход к формированию модели предметной области, основу которого составляют некоторые маркеры знаний (теги), заранее заданные учеными – участниками проекта Google Scholar Citations.

Следует отметить принципиальное отличие предложенной модели автоматического формирования модели предметной области от существующих, базирующихся на анализе текстовых корпусов (например, [3]) или непосредственном участии экспертов при выборе конкретных узлов и связей [1]. В данном случае эксперт-пользователь вкладывает лишь крупицы знаний в виде набора базовых тегов и небольших по объему словарей (десятки слов). В дальнейшем программа использует знания, заложенные самими авторами публикаций, теги отмеченные ими как главные. Т.е. экспертная середа в этом случае существенно расширяется.

Модель применена для отрасли науки «искусственный интеллект», но предложенный подход можно использовать и для других научных областей. Автором, в частности, построены подобные сети для направлений правовой науки и сложных сетей (Complex Networks).

Литература

[1] Добров Б.В., Соловьев В.Д., Лукашевич Н.В., Иванов В.В. Онтологии и тезаурусы. Модели, инструменты, приложения. Бином, 2009. – 173 с.

[2] Ландэ Д.В. Моделирование контентных сетей // Проблеми інформатизаціх та управління: Збірник наукових праць: Випуск 1(37). – К.: НАУ, 2012. – С. 78-84. (URL: http://dwl.kiev.ua/art/piu2012/)

[3] Ландэ Д.В., Снарский А.А.. Подход к созданию терминологических онтологий // Онтология проектирования, 2014. – № 2(12). – С. 83-91.

[4] Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.

[5] Чанышев О.Г. Автоматическое построение терминологической базы знаний // Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2008, Дубна, Россия, 2008. – С. 85-92.

[6] Kalogeraki V., Gunopulos D., Zeinalipour-Yazti D. A Local Search Mechanism for Peer-to-Peer Networks // Proc. of CIKM'02, McLean VA, USA, 2002.

A Domain Model Created on the Basis of Google Scholar Citations

Dmitry V. Lande

The technique of constructing terminological networks – domain models based on sensing informational networks is proposed. Such a network is considered a network of concepts relevant tag service Google Scholar Citations. Proposed approach can be applied in many areas of science.