УДК 001.8:004.7 doi: 10.18287/2223-9537-2015-5-3-328-335

ПОСТРОЕНИЕ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ ПУТЁМ ЗОНДИРОВАНИЯ СЕРВИСА GOOGLE SCHOLAR CITATIONS

Д.В. Ландэ

Институт проблем регистрации информации НАН Украины, Киев, Украина
dwlande@gmail.com

Аннотация

Предлагается алгоритм построения терминологических сетей – моделей предметных областей на основе зондирования большой информационной сети. В качестве такой сети рассматривается сеть понятий, соответствующих тегам сервиса Google Scholar Citations. Узлы в этой сети соответствуют понятиям, маркированным тегами, а ребра – некоторую семантическую связь между ними, определяемую смежными интересами отдельных авторов. Приведён специальный алгоритм сканирования ресурсов сервиса Google Scholar Citations для получения репрезентативного набора тегов как основы модели предметной области. На основе данной сети автоматически формируется релевантный список публикаций. Приведены правила построения списка библиографических ссылок. Предложенный подход может быть применён, в частности, к библиографическим базам данных, в которых в явном виде выделены авторы и как теги – ключевые слова. Данный подход можно применять для многих областей науки.

Ключевые слова: *модель предметной области, Google Scholar Citations, библиография, зондирование сети, визуализация сети.*

1 Задача создания модели предметной области

Сегодня под моделью предметной области (ПрО), в частности, понимают специальным образом сформированную сеть понятий, *онтологию*. Построение большой отраслевой онтологии, в частности, онтологии проектирования, – сложная научно-практическая проблема [1, 2]. Первый этап этого процесса – построение терминологической основы онтологии и определение семантических связей [3].

Задача автоматического создания таких сложных онтологий, как онтологии проектирования, требуют учитывать знания, изначально заложенные в некоторые тексты специалистами (учёными, экспертами). В качестве таких текстов могут рассматриваться специальные справочники, массивы документов [2], сетевых публикаций и т.п.

В работе представлен подход к созданию модели ПрО на основе зондирования большой информационной сети. В качестве такой сети рассматривается сеть понятий, которые отражаются в тегах[1] наукометрического сервиса *Google Scholar Citations*[2] (GSC). Именно эта сеть рассматривается как источник информации, используемой для построения сети понятий. На рисунке 1 приведён фрагмент интерфейса страницы сервиса GSC, соответствующий заданному заранее тегу multiagent_systems (многоагентные системы).

На интерфейсе, соответствующем данному тегу (label: multiagent_systems), постранично в ранжированном виде отображаются имена учёных, которые обозначили свои научные ин-

[1] В работе под тегом понимается обозначение понятия, научного направления, которое соответствует научным интересам учёного, и фиксируется либо самим учёным, либо экспертами.
[2] http://scholar.google.com/citations

тересы этим понятием, а также другими понятиями (например, для автора Nick Jennings определены ещё такие теги, как Artificial Intelligence, Computer Science, Intelligent Agents). Множество тегов-понятий образуют сеть, производную от биграфа[3] «учёный-понятия». Именно эту сеть будем рассматривать как некоторую модель ПрО. Узлы в этой сети соответствуют понятиям, маркированным тегами, а связи – некоторую семантическую связь между ними, определяемую смежными интересами отдельных авторов. Очевидно, эта связь может иметь вес, пропорциональный количеству авторов, которым приписывается соответствующая пара понятий.

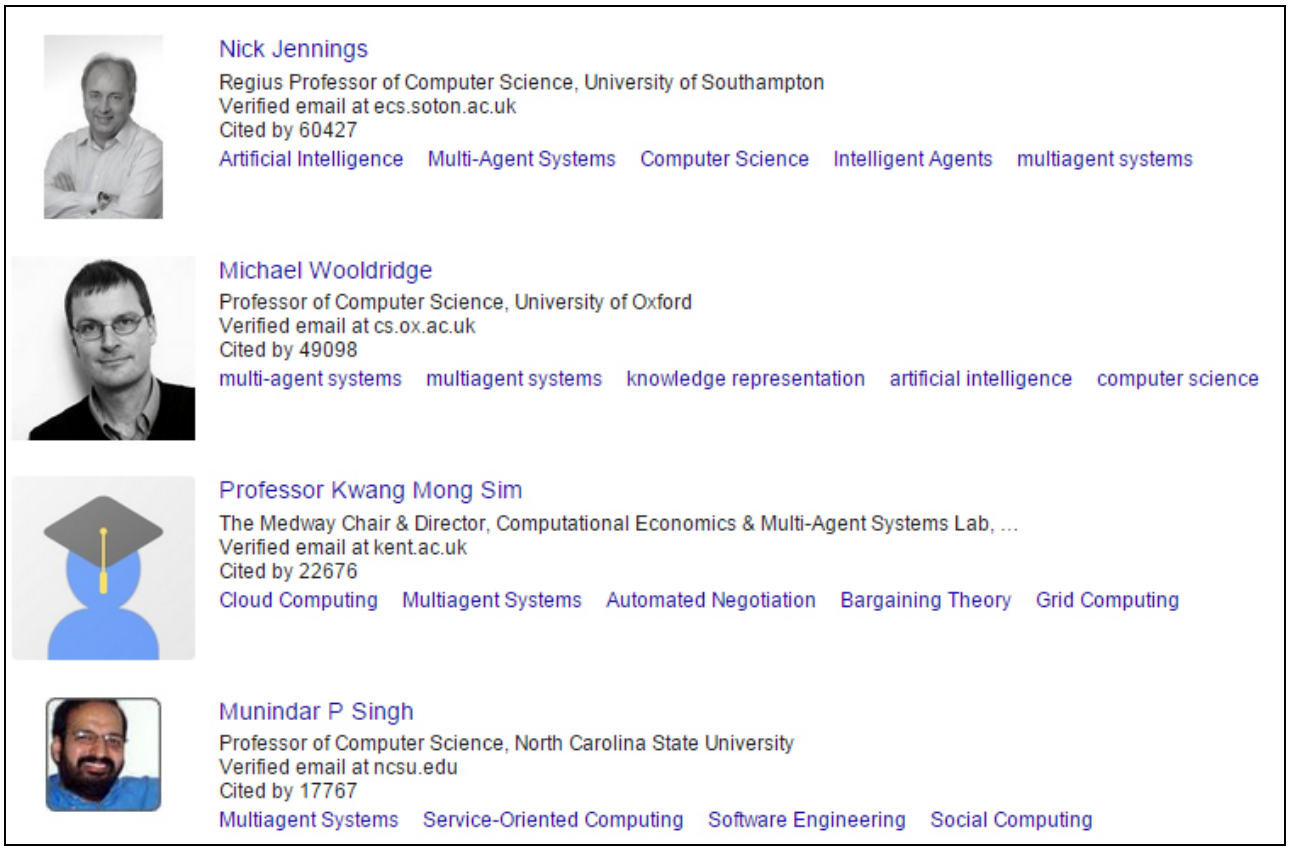


Рисунок 1 – Интерфейс страницы сервиса Google Scholar Citations

Конечно, теги, указанные отдельными учёными, могут относиться к различным отраслям науки. Однако предварительно проведённые исследования показывают, что на небольшой, но достаточно репрезентативной выборке (порядка сотни тегов), небольшая частота нетематических тегов обеспечивает их автоматическое «отсеивание».

Целью работы является описание подхода и алгоритмов автоматизированного формирования модели ПрО на примере направления многоагентных систем путём зондирования наукометрической сети. Для достижения этой цели автором разработан специальный алгоритм сканирования ресурсов сервиса GSC для получения репрезентативного набора тегов (обозначений понятий) как основы будущей модели ПрО. Под зондированием информационных сетей здесь понимается выборка небольшого объёма важнейшего содержания сетей, которые по технологическим причинам не подлежат полному сканированию.

[3] Биграф - это граф, множество вершин которого можно разбить на две части таким образом, что каждое ребро графа соединяет какую-то вершину из одной части с какой-то вершиной другой части. В рамках данной работы как одну часть можно рассматривать множество учёных, а другой – множество тегов.

2 Описание модели

Зондирование опорной модельной сети осуществляется по следующему алгоритму, в частности, применяемому при поиске ресурсов в пиринговых сетях[4] [4-6]:

1) выбирается определённое количество узлов опорной (зондируемой) сети, определяемых как базовые для новой сети, соответствующей результатам зондирования;
2) для каждого из рассматриваемых узлов опорной сети определяются смежные с ним узлы («соседи»), которые добавляются к создаваемой сети с результатами зондирования;
3) от текущего узла опорной сети осуществляется переход к соседнему узлу, имеющему наибольшую степень;
4) 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;
5) если перечень базовых узлов завершен, считается, что сеть, соответствующая результатам зондирования, построена.

Данный алгоритм проверялся для двух самых распространенных модельных сетей *Erdős-Rényi* (ER) и *Barabási-Albert* (BA) (рисунок 2) [6, 7]. 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 $p \times N \times (N - 1) / 2$ случайно выбранными связями.

а) б)

Рисунок 2 – Пример сети, построенной зондированием модельных сетей: (а) – Erdős-Rényi; (б) – Barabási-Albert

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 [7].

Автором изначально предполагалось, что сети понятий, естественным образом формируемые участниками сетевых сервисов, как и большинство информационных сетей обладают свойством безмасштабности [8, 9] (т.е. близкими по структуре к сети BA), что, однако, не всегда можно проверить, не имея всеобъемлющей информации. Если сеть такая масштабная, как, например, GSC, на помощь может прийти зондирование, в результате которого выполняется построение некоторой новой сети, лишь частично совпадающей с исходной. Отметим,

[4] Пи́ринговая или одноранговая, децентрализованная (англ. peer-to-peer, P2P — равный к равному) сеть – это оверлейная компьютерная сеть, основанная на равноправии участников.

что результаты любого зондирования не всегда верно отображают природу большой исследуемой сети – они во многом зависят именно от алгоритма процедуры зондирования. Вместе с тем, зондирование может служить базой для гипотез о структуре большой сети.

Визуально качественные результаты зондирования сетей ER и BA с близкими параметрами (1000 узлов и около 2000 связей) приведены на рисунке 2. Сравнение показывает, что связанные области (ветки), соответствующие отдельным понятиям в первом случае достаточно длинные, а узлов, по которым следует маршрут зондирования, в этом случае больше, чем во втором, более интересном для нас, случае. В рамках данного исследования важны именно качественные результаты, в частности, вид связанных цепочек, которыми моделируются ветки понятий. Поэтому изначально предусматривалось, что приведённый алгоритм при зондировании реальной сети будет быстро «зацикливаться» (и, соответственно, прерываться), что приведёт к ещё 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 GSC.

3 Зондирование сети Google Scholar Citations

Приведённый выше алгоритм, которой применялся к модельным сетям, был адаптирован к реальной сети тегов сервиса GSC следующим образом:

1) экспертным путём определяется небольшой перечень базовых тегов (ключевых слов, соответствующих наиболее важным понятиям);
2) выбирается тег из определённого экспертами перечня;
3) открываются страницы веб-сервиса, соответствующие этому тегу (максимальное количество таких страниц параметрически ограничивается заранее);
4) к создаваемой сети добавляются все теги, содержащиеся на выбранных страницах (соседние теги);
5) из соседних тегов выбирается тот, на страницы которого планируется перейти для дальнейшего анализа. Это тег с наибольшей степенью среди соседних тегов, который также удовлетворяет тематике выбранной ПрО и не входит в состав тех тегов, к страницам которых уже был осуществлен переход;
6) если такой тег выбран, то происходит переход к пункту 3;
7) если такого тега не существует, но перечень базовых тегов не завершён, 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 приведённым алгоритмом, процесс зондирования сети, начиная с определённого 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 учётом лексического состава тегов при полностью автоматическом сканировании. В случае автоматического выполнения алгоритма выполняется ограничение с помощью так называемых «плюс-» и «стоп-словарей»[5] – наборов специальных шаблонов. При этом само «зацикливание» является признаком перехода к следующему базовому тегу или завершению процесса зондирования.

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

[5] «Плюс-» и «стоп-словарь» в рамках данной работы – наборы шаблонов, подстрок, которые должны обязательно входить или, соответственно, не входить в строки, соответствующие тегам.

Для построения модели ПрО (в рассматриваемом примере для области многоагентных систем) экспертным путём были определены базовые теги на английском языке (здесь, в частности, различные написания, соответствующие одному и тому же понятию): *multiagent_system*; *multiagent_systems*; *multi_agent_systems*.

В качестве «плюс-словаря» использовался такой набор шаблонов: *multiagent*; *agent*; *artificial_intelligent*; *decentral*.

В рассматриваемом частном случае «стоп-словарю» соответствовал набор шаблонов (исключалась медико-биологическая тематика): *bio*; *medic*; *health*.

На рисунке 3 приведён пример - центральный фрагмент сети понятий ПрО, построенной в соответствии с приведённым алгоритмом по указанным базовым тегам.

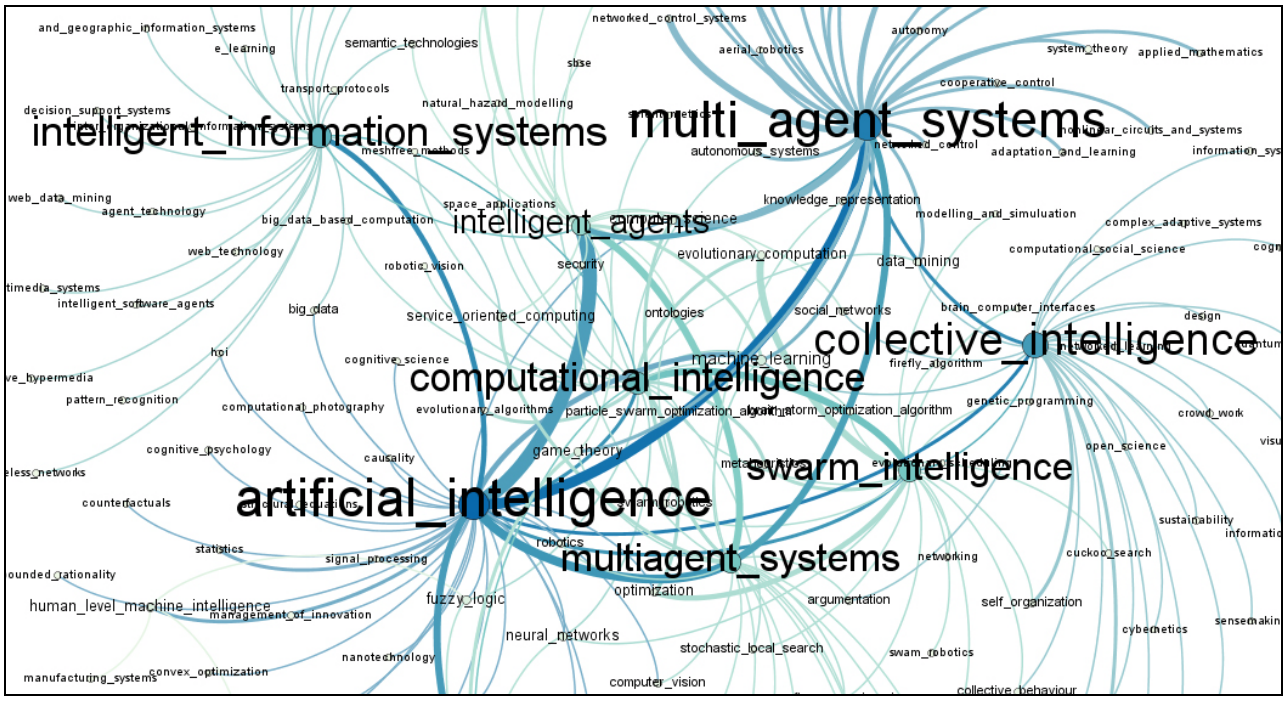


Рисунок 3 – Центральный фрагмент построенной сети понятий

При приведённых выше трёх базовых тегах общее количество узлов-тегов, которые были охвачены алгоритмом, составило 165, а количество нетерминальных узлов – лишь 12.

4 Построение библиографических списков

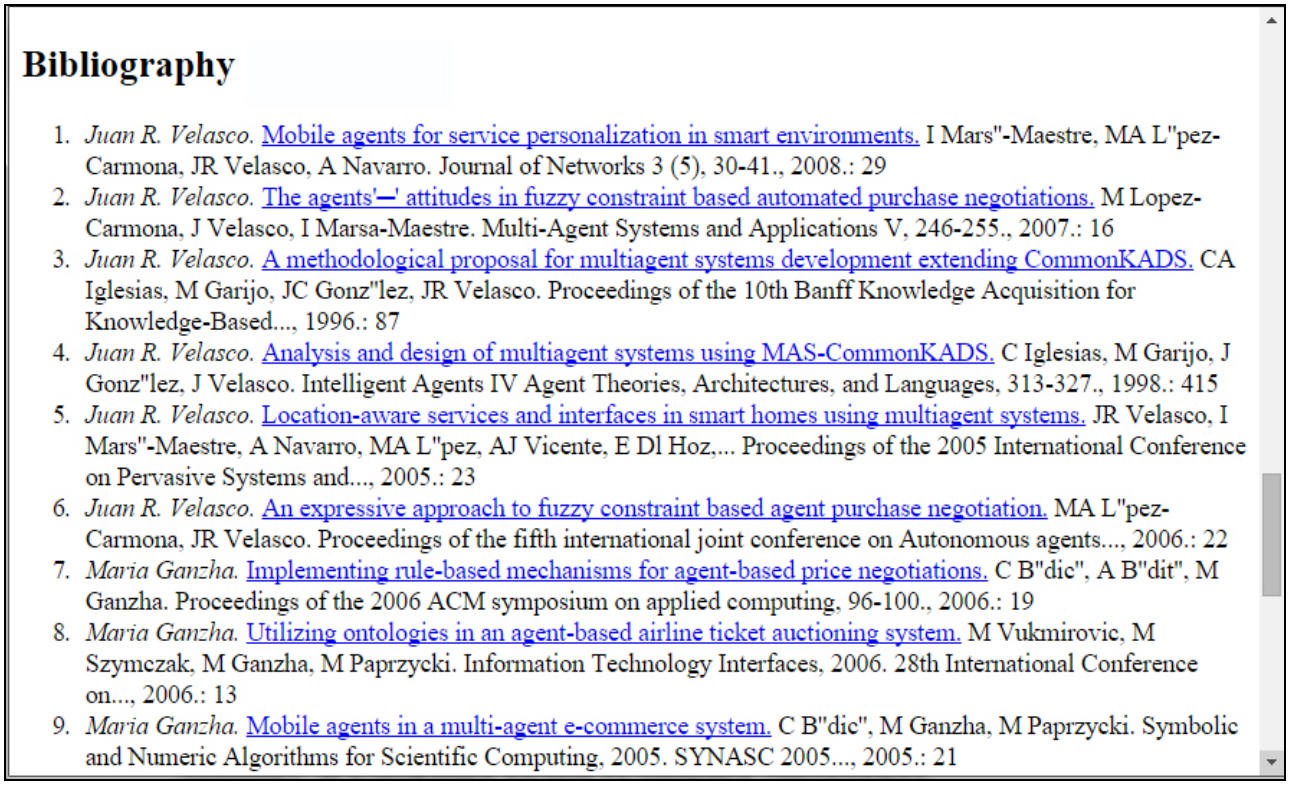
На основании построенной в результате зондирования сервиса GSC сети и возможностей данного сервиса могут автоматически составляться библиографические списки, соответствующие наиболее цитируемым работам в выбранной области за указанные промежутки времени.

Правила построения библиографических ссылок достаточно простые и охватывают такие шаги:

1) последовательно выбирается заранее заданное количество наиболее весомых узлов построенной сети;
2) в соответствии с этими узлами-тегами формируются запросы к сервису GSC и отбираются наиболее цитируемые авторы за указанный период времени;
3) открываются страницы выбранных авторов, на которых приводятся ссылки на публикации, с указанием их цитируемости;

4) среди наиболее цитируемых работ автора выбираются публикации, заголовки которых соответствуют «плюс-» и «стоп-словарям» (примеры приведены выше);
5) в случае необходимости из отобранных публикаций выбираются только те, которые содержат ссылки на полные тексты в формате PDF.

На рисунке 4 приведён пример автоматически сформированного библиографического списка со ссылками на PDF-файлы публикаций.



Bibliography

1. *Juan R. Velasco.* Mobile agents for service personalization in smart environments. I Mars"-Maestre, MA L"pez-Carmona, JR Velasco, A Navarro. Journal of Networks 3 (5), 30-41., 2008.: 29
2. *Juan R. Velasco.* The agents'—' attitudes in fuzzy constraint based automated purchase negotiations. M Lopez-Carmona, J Velasco, I Marsa-Maestre. Multi-Agent Systems and Applications V, 246-255., 2007.: 16
3. *Juan R. Velasco.* A methodological proposal for multiagent systems development extending CommonKADS. CA Iglesias, M Garijo, JC Gonz"lez, JR Velasco. Proceedings of the 10th Banff Knowledge Acquisition for Knowledge-Based..., 1996.: 87
4. *Juan R. Velasco.* Analysis and design of multiagent systems using MAS-CommonKADS. C Iglesias, M Garijo, J Gonz"lez, J Velasco. Intelligent Agents IV Agent Theories, Architectures, and Languages, 313-327., 1998.: 415
5. *Juan R. Velasco.* Location-aware services and interfaces in smart homes using multiagent systems. JR Velasco, I Mars"-Maestre, A Navarro, MA L"pez, AJ Vicente, E Dl Hoz,... Proceedings of the 2005 International Conference on Pervasive Systems and..., 2005.: 23
6. *Juan R. Velasco.* An expressive approach to fuzzy constraint based agent purchase negotiation. MA L"pez-Carmona, JR Velasco. Proceedings of the fifth international joint conference on Autonomous agents..., 2006.: 22
7. *Maria Ganzha.* Implementing rule-based mechanisms for agent-based price negotiations. C B"dic", A B"dit", M Ganzha. Proceedings of the 2006 ACM symposium on applied computing, 96-100., 2006.: 19
8. *Maria Ganzha.* Utilizing ontologies in an agent-based airline ticket auctioning system. M Vukmirovic, M Szymczak, M Ganzha, M Paprzycki. Information Technology Interfaces, 2006. 28th International Conference on..., 2006.: 13
9. *Maria Ganzha.* Mobile agents in a multi-agent e-commerce system. C B"dic", M Ganzha, M Paprzycki. Symbolic and Numeric Algorithms for Scientific Computing, 2005. SYNASC 2005..., 2005.: 21

Рисунок 4 – Фрагмент библиографического списка со ссылками на PDF-файлы публикаций

Заключение

В предложенной модели ПрО онтологические связи понимаются как связи между областями интересов отдельных учёных. Фактически рассматривается компактификация биграфа «учёный – научные понятия, его интересующие».

Предложен и реализован подход к формированию модели ПрО, основу которого составляют некоторые маркеры понятий (теги), заранее заданные учёными (или, в редких случаях, приписываемые учёным) – участниками проекта Google Scholar Citations.

Следует отметить принципиальное отличие предложенной модели автоматического формирования модели ПрО от существующих, базирующихся на анализе текстовых корпусов (например, [2]) или непосредственном участии экспертов при выборе конкретных узлов и связей [1]. Здесь эксперт-пользователь вкладывает лишь крупицы знаний в виде набора базовых тегов и небольших по объёму словарей тегов и шаблонов. В дальнейшем программа использует знания, заложенные самими авторами публикаций, теги, отмеченные ими как главные. Т.е. экспертная среда в этом случае существенно расширяется.

Реализован алгоритм, в соответствии с которым на основании построенной сети формируется библиографический список наиболее цитируемых работ в данной ПрО, представленных в базе данных сервиса Google Scholar Citations.

Подобный подход может быть применён, в частности, к библиографическим базам данных, в которых в явном виде выделены авторы и как теги – ключевые слова.

Модель применена для отрасли науки «многоагентные системы», но предложенный подход можно использовать и для других научных областей. Автором, в частности, построены подобные сети для направлений искусственного интеллекта, глубинного анализа текстов (Text Mining) и сложных сетей (Complex Networks).

Благодарности

Статья обобщает некоторые результаты, полученные автором в рамках выполнения научно-исследовательской работы НАН Украины «Разработка теоретических основ моделирования информационных сетей на основе методологии информационного поиска» (Гипернет-2013).

Автор благодарен своим коллегам д.т.н. А.Г. Додонову, д.ф.-м.н. А.А. Снарскому и к.т.н. В.Г. Путятину за обсуждение и конструктивные предложения, относящиеся к методам исследования и результатам, представленным в статье.

Список источников

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

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

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

[4] ***Zeinalipour-Yazti, D.*** Information Retrieval in Peer-to-Peer Networks / D. Zeinalipour-Yazti, V. Kalogeraki, D. Gunopulos // IEEE CiSE Magazine. Special Issue on Web Engineering. 2004. – P. 1-13.

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

[6] ***Yang, B.*** Efficient Search in Peer-to-Peer Networks / B. Yang, H. Garcia-Molina // Proc. of ICDCS'02. Vienna, Austria, 2002.

[7] ***Erdős, P.*** On The Evolution of Random Graphs / P. Erdős, A. Rényi // Magyar Tud. Akad. Mat. Kutató Int. Közl. 5, 1960. – P. 17-61.

[8] ***Réka, A.*** Statistical mechanics of complex networks / A. Réka, A.-L. Barabási // Reviews of Modern Physics 74, 2002. – P. 47–97.

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

CREATION OF A DOMAIN MODEL BY PROBING GOOGLE SCHOLAR CITATIONS

D.V. Lande

Institute for Information Recording of NAS of Ukraine, Kiev, Ukraine
dwlande@gmail.com

Abstract

The algorithm of constructing terminological networks – domain models based on sensing informational networks is proposed. A network of the concepts relevant to tags of the Google Scholar Citations service is considered to be a terminological network. Nodes in this network correspond to the concepts marked by tags. Edges correspond to some semantic link between them determined by adjacent interests of certain authors. The special algorithm of scanning the

resources of the Google Scholar Citations service for receiving a representative set of tags as domain models bases is given. A relevant list of publications is automatically formed based on this network. Rules of creation of the list of bibliographic links are provided. The offered approach can be applied, in particular, to bibliographic databases in which authors and as tags – keywords are allocated in an explicit form. Proposed approach can be applied to many areas of science.
Key words: *Domain model, Google Scholar Citations, Bibliography, Network probing, Network visualization.*

Acknowledgment

The article summarizes the results of the authors research, obtained while working on a "Development of theoretical principles of modeling of information networks based on the methodology of information retrieval" (Hypernet-2013) – a Ukrainian National Academy of Science project.

The author is grateful to his colleagues dr. A.G. Dodonov, dr. A.A. Snarsky and V.G. Putiain for discussions and productive feedback on methods of the research and the results, presented in the paper.

References

[1] ***Dobrov, B.V.*** Ontologii i tezaurusy. Modeli, instrumenty, prilozhenija [Ontologies and thesauri. Models, instruments, applications] / B.V. Dobrov, V.D. Solov'ev, N.V. Lukashevitsh, V.V. Ivanov. – Moscow: Binom, 2009. – 173 p. (In Russian).

[2] ***Lande, D.V.*** Podhod k sozdaniy terminologicheskih ontologii [Approach to the creation of terminological ontologies] / D.V. Lande, A.A. Snarski // Ontologija proektirovanija. 2014. №2(12). – P. 83-91. (In Russian).

[3] ***Chanishev, O.G.*** Avtomaticheskoe postroenie terminologicheskoj bazy znanij [Automatic generation of terminological knowledge base] / O.G. Chanishev // Proceedings of the 10th Russian scientific conference «E-libraries: perspective methods, relevent collections» – RCDL'2008, Dubna, Russia, 2008. – P. 85-92. (In Russian).

[4] ***Zeinalipour-Yazti, D.*** Information Retrieval in Peer-to-Peer Networks / D. Zeinalipour-Yazti, V. Kalogeraki, D. Gunopulos // IEEE CiSE Magazine. Special Issue on Web Engineering. 2004. – P. 1-13.

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

[6] ***Yang, B.*** Efficient Search in Peer-to-Peer Networks / B. Yang, H. Garcia-Molina // Proc. of ICDCS'02. Vienna, Austria, 2002.

[7] ***Erdős, P.*** On The Evolution of Random Graphs / P. Erdős, A. Rényi // Magyar Tud. Akad. Mat. Kutató Int. Közl. 5, 1960. – P. 17-61.

[8] ***Réka, A.*** Statistical mechanics of complex networks / A. Réka, A.-L. Barabási // Reviews of Modern Physics 74, 2002. – P. 47–97.

[9] ***Lande, D.V.*** Modelirovanie kontentnyh setej [Content network modelling]/ D.V. Lande // Problemy informatizacich ta upravlinnja: Zbirnyk nauobych prac': Vypusk 1(37). – Kiev: NAU, 2012. – P 78-84. (http://dwl.kiev.ua/art/piu2012/) (In Ukrainian).

Сведения об авторе

Ландэ Дмитрий Владимирович, 1959 г. рождения. Окончил Киевский государственный университет им. Т.Г. Шевченко, механико-математический факультет в 1981 г., доктор технических наук (2006). Заведующий отделом специализированных средств моделирования Института проблем регистрации информации НАН Украины, профессор Национального технического университета «Киевский политехнический институт», академик Украинской академии наук (УАН), член Российской ассоциации искусственного интеллекта. В списке научных трудов более 300 работ в области информационного поиска, динамики информационных потоков, информационных сетей.

Dmitry Vladimirovich Lande (b. 1959) graduated from the Shevchenko Kiev State University, mechanics and mathematics faculty in 1981, Dr. of Sciences (2006). He is department head of the Institute for Information Recording of NAS of Ukraine, professor at National Technical University of Ukraine "Kiev Politechnical Institute". He is full member of the Ukrainian Academy of Sciences (UAS), Russian Association for Artificial Intelligence (RAAI) member. He is co-author of over 300 scientific articles, books and abstracts in the field of information retrieval and information flows dynamics.