INFORMATION TECNOLOGY

УДК 004.67

ДМИТРИЙ ЛАНДЕ

ПОСТРОЕНИЕ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ПУТЕМ ЗОНДИРОВАНИЯ СЕРВИСА GOOGLE SCHOLAR CITATIONS

Предлагается методика построения терминологических сетей – моделей предметных областей на основе зондирования данных сетевого информационного сервиса. Как такой сервис рассматривается Google Scholar Citations. Данный подход применяется для предметной области информационной безопасности, но может быть применен для многих других областей науки и технологий.

Ключевые слова: предметная область, модель предметной области, информационная сеть, зондирование сети, сеть понятий.

Постановка проблемы. Под моделью предметной области понимают специальным образом сформированную сеть понятий (онтологию). Построение большой отраслевой онтологии – сложная и актуальная проблема. Первый этап решения этой проблемы – определение терминологической основы и определение семантических связей [1, 2].

В этой работе представляется подход к созданию модели предметной области (информационная безопасность) на основе зондирования информационной сети крупного информационного сервиса. 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 information security (информационная безопасность). На интерфейсе, соответствующем данному тегу (label: information_security) постранично в ранжированном виде отображаются имена ученых, которые отметили свою деятельность этим тегом, а также другие теги, приписанные ими (например, Moti Yung определил для себя еще такие теги: cryptography, information security, computer security, distributed algorithms, privacy). Множество тегов образуют сеть, производную от биграфа «ученный–теги». Эту сеть и предлагается рассматривать как некоторую онтологическую модель предметной области [3]. Узлы в этой сети соответствуют понятиям, маркированным тегами, а связи – некоторой семантической связи между ними.

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.

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).

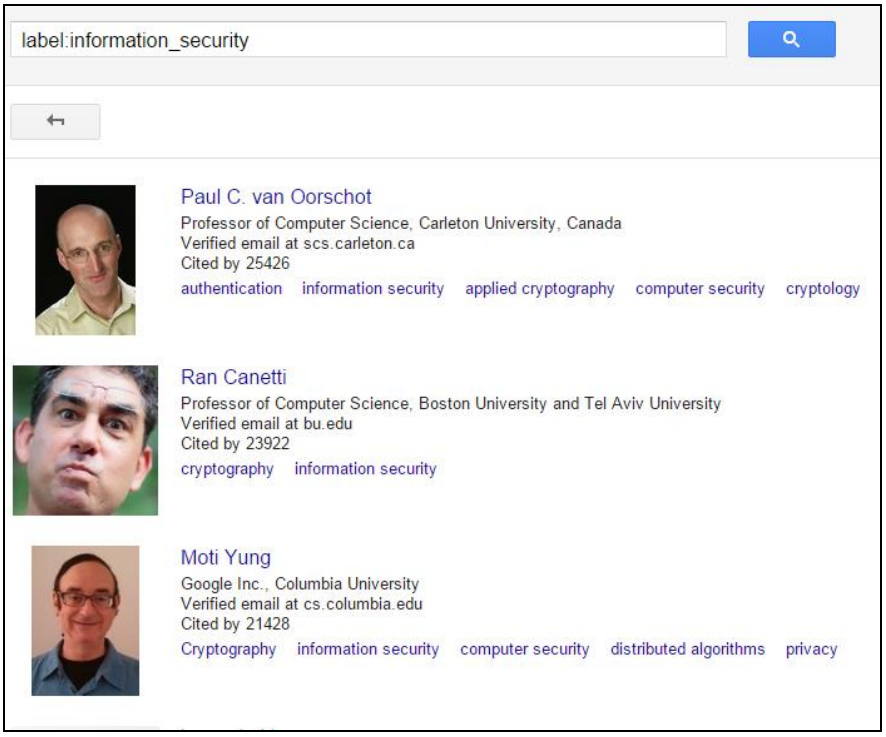

Рисунок 1 – Интерфейс страницы сервиса Google Scholar Citations

Для поиска необходимых данных в таких сетях применяется несколько моделей [4, 5]. В модели «широкого первичного поиска» (Breadth First Search, BFS) запрос из некоторого стартового узла адресуется ко всем своим 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 случае успешного поиска формируется сообщение-отзыв (QueryHit), которое включает информацию о релевантных запросу узлах, 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 учета контента узлов-соседей, минимизации количества сообщений между узлами и количества узлов, опрашиваемых для каждого поискового запроса [4]. В этом случае 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:

1. Выбирается определенное 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 (в простейшем случае как аналог наиболее релевантного узла).
4. Если имеет место «зацикливание» (выбирается узел, к которому уже был осуществлен переход), происходит переход к следующему по степени соседнему узлу. Если таких узлов не осталось – осуществляется выбор следующего базового узла и переход к пункту 2.
5. 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) [5, 6]. 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.

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.

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. Автором изначально предполагалось, что сети понятий, естественным образом формируемые участниками сетевых сервисов тоже обладают свойством безмасштабности, что не всегда можно проверить, не имея всеобъемлющей информации. Если сеть такая сложная и большая, как, например, Google Scholar Citations, на помощь может прийти зондирование. Отметим, однако, что результаты любого зондирования не всегда верно отображают природу большой исследуемой сети – они во многом зависят именно от алгоритма.

Визуально качественные результаты зондирования сетей ER и BA с близкими параметрами (1000 узлов, около 2000 связей) показывают, 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, случае. 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.

Зондирование сети Google Scholar Citations. Алгоритм, примененный к модельным сетям, был адаптирован к реальной сети тегов сервиса Google Scholar Citations следующим образом:

1. Экспертным путем определяется небольшой перечень базовых тегов (ключевых слов, соответствующих наиболее важным понятиям).
2. Выбирается тег из определенного 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. Этот тег с наибольшей степенью среди соседних тегов, который также удовлетворяет тематике выбранной предметной области (выбор осуществляется экспертами или автоматически по шаблонам слов, соответствующих названиям тематических тегов) 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 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.

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 информационной безопасности) экспертным путем было определено 10 базовых тегов на английском языке: information_security; cryptography; computer_security; cryptology; network_security; traffic_analysis; internet_security; cybersecurity; big_data; system_security.

В качестве шаблонов слов для выбора тематических тегов с целью автоматических «переходов» в процедуре зондирования использовались такие шаблоны: _informat; _secur; ypt; _computer; _embedded; _system; _privac; _applied; _network; _trusted; _anonym; _traffic; _analys; _internet; _data; _privac; _cyber; _managem; _identity; _mashine; _agent.

На рис. 3 представлен общий вид сети понятий предметной области, построенной в соответствии с приведенным алгоритмом по указанными базовым тегам, а на рис. 4 фрагмент с отмеченными тегами.

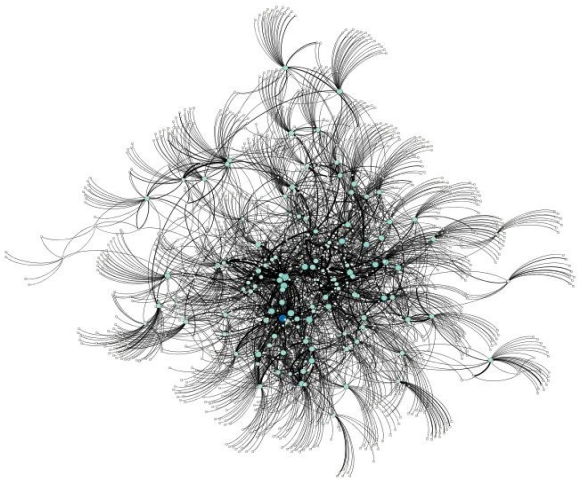

Рисунок 3 – Общий вид сети понятий предметной области

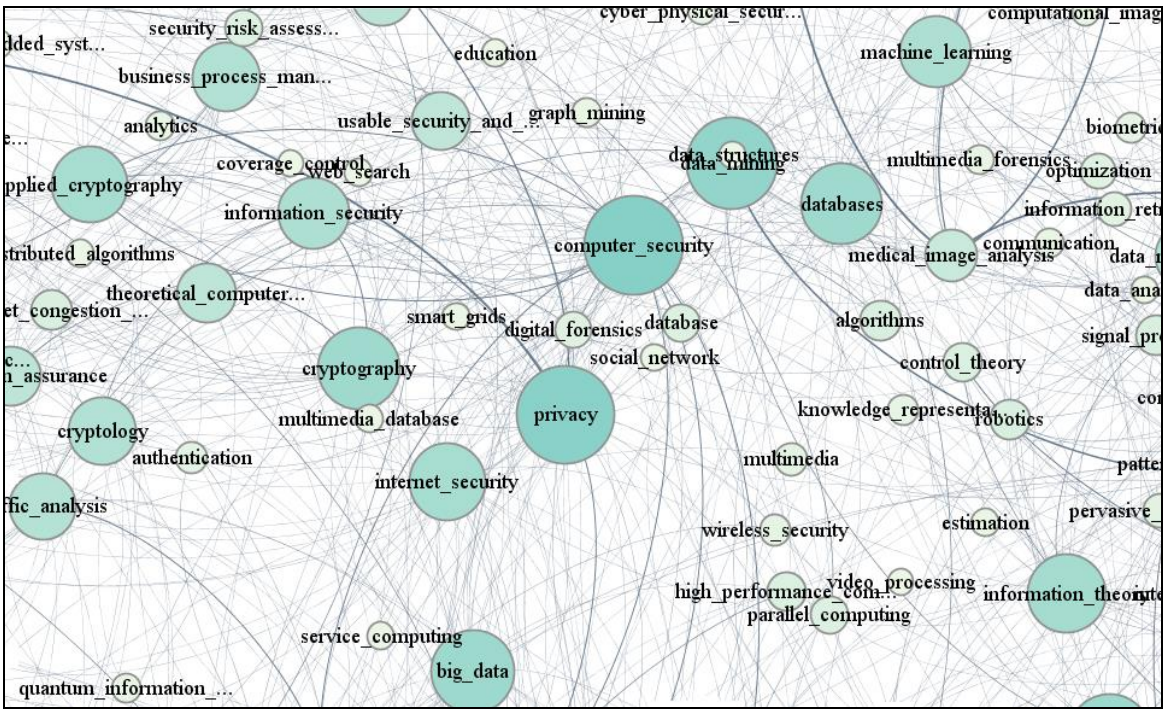

Рисунок 4 – Центральный фрагмент сети понятий

Построенная сеть понятий оказалась связной. При количестве базовых тегов 10, общее количество узлов-тегов, которые были охвачены алгоритмом, составили 1451, а количество нетерминальных узлов – лишь 138. Распределение степеней этих узлов, приведенное на рис. 5, 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 10.

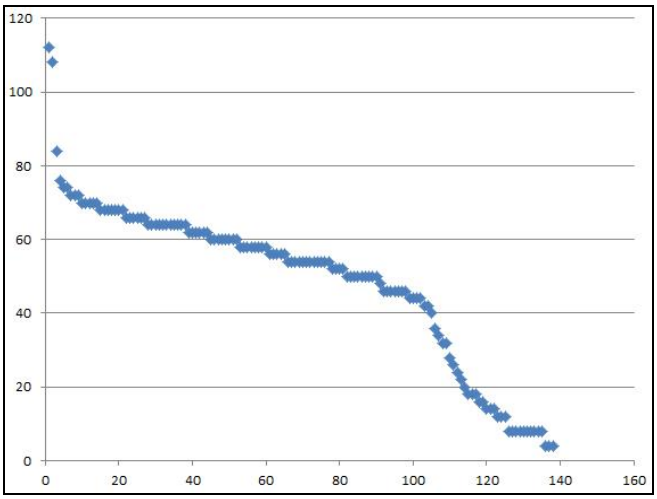

Рисунок 5 – Распределение степеней узлов-тегов сети понятий

Ранжирование узлов в построенной сети зондирования возможно по свойствам, обуславливаемым сетевой структурой, ссылками («направлением движения» при зондировании). Например, для определения авторитетности узла как слова – источника порождения словосочетаний или как составного термина, состоящего из отдельных важных слов, можно анализировать сеть, выбирая при этом наиболее важных «авторов» или «хабов». Для решения этой задачи предлагается использовать известный алгоритм ранжирования веб-страниц, основанных на связях, HITS (Hyperlink Induced Topic Search), предложенный Дж. Клейнбергом [7], который может применяться, например, наряду с алгоритмами PageRank (в этом случае – оценка единственная, интегрированная) или Salsa.

Алгоритм HITS обеспечивает выбор из информационного массива лучших «авторов» (узлов, на которые введут ссылки) и «посредников» (узлов, от которых идут ссылки включения). В рассматриваемом случае термин является хорошим посредником, если от него идут связи на важные словосочетания, и наоборот, термин (словосочетание) является хорошим автором, если на него ведут связи от важных авторов. В соответствии с алгоритмом HITS для каждого узла сети v_j рекурсивно вычисляется его значимость как автора $a(v_j)$ и посредника $h(v_j)$ по формулам:

$$a(v_j) = \sum_{i \to j} h(v_i);\ h(v_j) = \sum_{j \to i} a(v_i).$$

В данных формулах суммирование производится по всем узлам, которые ссылаются (или на которые ссылаются – во второй формуле) на данный узел.

В табл. 1 и 2 приведены списки тегов являющихся лучшими авторами и тегами в рамках предметной области «информационная безопасность» по построенной сети зондирования (модели предметной области).

Наиболее интересными с семантической точки зрения в рассматриваемой сети оказались узлы с наибольшим значением авторства и посредничества (security; privacy; wireless_networking; networking; computer_security; network_management; computer_networks; network_security).

Таблица 1 – 20 наиболее значимых «автора» сети зондирования

1	**security**
2	**networking**
3	distributed_systems
4	**wireless_networks**
5	**computer_networks**
6	**privacy**
7	**computer_security**
8	**network_security**
9	cryptography
10	operating_systems
11	information_security
12	data_mining
13	**network_management**
14	computer_architecture
15	big_data
16	networks
17	information_theory
18	wireless_sensor_networks
19	databases
20	computer_networking

Таблица 2 – 20 наиболее значимых «посредников» сети зондирования

1	internet_security
2	**security**
3	traffic_analysis
4	**privacy**
5	**wireless_networking**
6	**networking**
7	operating_systems
8	internet_measurement
9	network_measurement
10	cybersecurity
11	**computer_security**
12	**network_management**
13	network_measurements
14	applied_cryptography
15	system_security
16	internet_routing
17	systems
18	**computer_networks**
19	**network_security**
20	distributed_systems

Выводы:

1. В предложенной модели предметной области как онтологические связи применяются связи между областями интересов отдельных ученых. Фактически рассматривается компактификация биграфа «ученый – области науки и технологий, его интересующие».

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

3. Предложено использование алгоритма HITS для выбора наиболее важных понятий из предметной области в сети зондирования.

Модель применена для отрасли науки и технологий «информационная безопасность», но предложенный подход можно использовать и для других областей. Автором, в частности, построены подобные сети для направлений правовой науки и искусственного интеллекта.

СПИСОК ИСПОЛЬЗОВАНОЙ ЛИТЕРАТУРЫ

1. Ланде Д. В. Елементи комп'ютерної лінгвістики в правовій інформатиці / Д. В. Ланде. – К. : НДІІП НАПрН України, 2014. – 168 с.

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

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

4. 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. – P. 300-307.

5. Ландэ Д. В. Интернетика : навигация в сложных сетях : модели и алгоритмы / Д. В. Ландэ, А. А. Снарский, И. В. Безсуднов. – М. : Либроком (Editorial URSS), 2009. – 264 с.

6. Ландэ Д. В. Моделирование контентных сетей / Д. В. Ланде // Проблеми інформатизації та управління : збірник наукових праць. – К. : НАУ, 2012. – Вип. 1 (37). – С. 78–84.

7. Kleinberg J. Authoritative sources in a hyperlinked environment / J. Kleinberg // In Processing of ACM-SIAM Symposium on Discrete Algorithms. – 1998. – № 46 (5). – P. 604–632.

Статья поступила в редакцию 30.01.2015.

REFERENCE

1. Lande, D. V. (2014), *Elements of Computational Linguistics in legal informatics* [*Elementy kompiuternoi linhvistyky v pravovii informatytsi*], NDIIP NAPrN, Kiev, 168 p.

2. Dobrov, B. V., Solovyov, V. D., Lukashevich, N. V., Ivanov, V. V. (2009), *Ontologies and thesauri: models, tools, applications* [*Ontologii i tezaurusy: modeli, instrumenty, prilozheniia*], Binom, Moscow, 173 p.

3. Lande, D. V., Snarskii, A. A. (2014), *The approach to the creation of terminological ontologies* [*Podkhod k sozdaniiu terminologicheskikh ontologii*], Ontology engineering, No. 2 (12), pp. 83-91.

4. Kalogeraki, V., Gunopulos, D., Zeinalipour-Yazti, D. (2002), *A Local Search Mechanism for Peer-to-Peer Networks*, Proc. of CIKM'02, McLean VA, USA, pp. 300-307.

5. Lande, D. V., Snarskii, A. A., Bezsudnov, I. V. (2014), *Internetika: Navigation in complex networks: models and algorithms* [*Internetika : navigatciia v slozhnykh setiakh : modeli i algoritmy*], Librokom (Editorial URSS), Moscow, 264 p.

6. Lande, D. V. (2012), *Simulation of contact networks* [*Modelirovanie kontentnykh setei*], Problems of Informatization and management: research papers collection, NAU, No. 1 (37), pp. 78-84.

7. Kleinberg, J. (1998), *Authoritative sources in a hyperlinked environment*, In Processing of ACM-SIAM Symposium on Discrete Algorithms, No. 46 (5), pp. 604-632.

ДМИТРО ЛАНДЕ

ПОБУДОВА МОДЕЛІ ПРЕДМЕТНОЇ ОБЛАСТІ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ ШЛЯХОМ ЗОНДУВАННЯ СЕРВІСУ GOOGLE SCHOLAR CITATIONS

Пропонується методика побудови термінологічних мереж – моделей предметних областей на основі зондування даних мережевого інформаційного сервісу. Як такий сервіс розглядається Google Scholar Citations. Даний підхід використовується для предметної

області інформаційної безпеки. Разом з тим, він придатний до застосування в багатьох інших областях науки та технологій.

Ключові слова: предметна область, модель предметної області, інформаційна мережа, зондування мережі, мережа понять.

DMITRII LANDE

CONSTRUCTION DOMAIN MODEL OF INFORMATION SECURITY BY PROBING GOOGLE SCHOLAR CITATIONS SERVICE

The technique of constructing terminological networks – domain models based on the sensing data of the network information service – proposed. As this service is considered Google Scholar Citations. This approach is applied to the subject area of information security, but can be applied to many other areas of science and technology.

Keywords: subject area, domain model, network probing, information network, concepts network.

Дмитрий Владимирович Ланде, доктор технических наук, старший научный сотрудник, заведующий отделом специализированных средств моделирования, Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина.
E-mail: dwlande@gmail.com.

Дмитро Володимирович Ланде, доктор технічних наук, старший науковий співробітник, завідувач відділом спеціалізованих засобів моделювання, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.

Dmitrii Lande, doctor of technical science, senior researcher, head of the specialized modeling tools department, Institute of problems of information registration of National academy of science of Ukraine, Kyiv, Ukraine.

УДК 004.78; 004.891.2; 007.3

ОЛЕКСІЙ КОВАЛЕНКО

ЗАСТОСУВАННЯ ХМАРНИХ ТЕХНОЛОГІЙ ПРИ ПОБУДОВІ СИСТЕМ СИТУАЦІЙНОГО УПРАВЛІННЯ

Розглядаються питання побудови та організації багатопрофільних систем ситуаційного управління, побудованих на принципах хмарних обчислень. Аналізуються особливості створення та ефективного використання таких систем. Запропоновано класифікацію корпоративних додатків з точки зору їх функціональності.

Ключові слова: системи ситуаційного управління, інформаційні сервіси, хмарні обчислення.

Особливості створення сучасних корпоративних систем. Розробка сучасних корпоративних інформаційних систем характеризується комплексним характером і різною направленістю вирішуваних завдань. Це висуває особливі вимоги до підрозділів, які займаються вирішенням завдань інформатизації на підприємстві – створенням, впровадженням, підтримкою, адмініструванням і модернізацією.