OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822: 004.912

МОДЕЛИ ПРЕДМЕТНЫХ ОБЛАСТЕЙ В СИСТЕМАХ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ МОНИТОРИНГА ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

Додонов А.Г., Ландэ Д.В., Коваленко Т.В.

Институт проблем регистрации информации НАН Украины,
г. Киев, Украина

dodonov@ipri.kiev.ua

dwlande@gmail.com

2005ste@ukr.net

Описываются подходы к построению моделей предметных областей в системах поддержки принятия решений. Как источники знаний рассматриваются текстовые корпусы и ресурсы современных наукометрических сервисов. Представлена архитектура системы информационной поддержки принятия решений, идеология создания и использования онтологий для построения сюжетов информационной поддержки принятия решений. Предложена методология аналитического исследования, которая базируется на использовании инструментальных средств анализа и визуализации информационных потоков и сетевых структур.
Ключевые слова: поддержка принятия решений, онтология, модель предметной области, информационное пространство, контент-мониторинг.

Введение

В настоящее время все большее применение находят знание-ориентированные информационные системы с онтолого-управляемой архитектурой [Палагин, 2012], которые обеспечивают компьютерную обработку естественно-языковых объектов, описывающих предметные знания: объекты, их свойства и понятия предметной области [Ланде, 2013], [Добров, 2009].

В частности, системы с онтолого-управляемой архитектурой находят свое применение в системах информационной поддержки процессов противодействия информационным кампаниям, деструктивным внешним и внутренним информационным воздействиям, что является актуальной проблемой современности [Горбулін, 2009], [Шульц, 2011].

В качестве основных задач такой системы рассмотрим:

- построение сценариев противодействия информационным деструктивным воздействиям на основе некоторой онтологии понятий;
- контент-мониторинг (непрерывный содержательный анализ) информационного пространства на основе знаний экспертов;
- выявление закономерностей (трендов) и аномалий путем анализа динамики изменения значений отдельных факторов;
- выявление информационных воздействий и информационных операций;
- прогнозирование развития информационных сюжетов и ситуаций;
- оценка эффективности процедур поддержки принятия решений.

Соответственно, для реализации такой системы информационной поддержки процессов, связанных, в частности, национальной безопасностью необходимо:

- создать онтологию понятий предметной области (узлов – факторов безопасности и соответствующих причинно-следственных (каузальных) связей – зависимости факторов), определить вид целевой функции безопасности объектов-узлов этой онтологии в зависимости от значений факторов безопасности;
- постоянно актуализировать значение факторов безопасности и связей в зависимости от результатов мониторинга информационного пространства и знаний экспертов;
- определять возможные сценарии на основе анализа онтологии и выявления соответствующих частных онтологий;

- анализировать динамику изменения значений отдельных факторов и связей с целью выявления закономерностей, прогнозирования;
- постоянно проводить оценку эффективности проводимой информационной поддержки.

Для решения приведенных выше задач системы информационной поддержки принятия решений, предполагается, что система информационной поддержки должна состоять из трех основных подсистем (рис. 1): подсистемы создания онтологии понятий, подсистемы мониторинга информационного пространства, подсистемы аналитической обработки, и соответствующих интерфейсов с администраторами и пользователями.

Рисунок 1 – Архитектура системы информационной поддержки

1. Задача формирования онтологии

Онтологический подход логично интегрируется со сложными динамическими системами. Вместе с тем, сами онтологии, представляемые в виде семантических сетей, когнитивных карт, каузальных сетей и т.п., являются динамическими объектами, состав их узлов и ребер постоянно изменяется, причем эксперты не всегда в состоянии оперативно отслеживать эти изменения. Таким образом, современная онтологическая система, необходимая для принятия решений, должна содержать характеристики как предметной области, так и внешней информационной среды, с которой система должна взаимодействовать.

Онтология в данном случае представляет собой функциональный аналог базы знаний, отражающей знания экспертов о предметной области, т.е. в качестве узлов графа онтологии выбираются важнейшие факторы предметной области обеспечения безопасности, а в качестве связей - причинно-следственные связи между факторами (с математической точки зрения – граф с направленными ребрами) [Шульц, 2014]. Узлам и связям приписываются числовые значения, которые в дальнейшем могут корректироваться. Связи также могут иметь разный вес (силу воздействия) и быть как положительными (увеличение значения первого фактора приводит к увеличению значения второго фактора), так и отрицательными (увеличение значения первого фактора приводит к уменьшению значения второго фактора).

При формировании предметных онтологий, применяемых в системах поддержки принятия решений, решается несколько содержательных задач, среди которых:

- выявление узлов – основных понятий и соответствующих им слов и словосочетаний из соответствующей предметной области;
- выявление различных семантических связей между узлами и соответствующих им понятий;
- ранжирование понятий, выявление главных понятий в предметной области;
- ранжирование связей, выявление главных связей.

Построение онтологии в процессе изучения некоторой системы может рассматриваться как прямая задача, где заранее известна схема управления, основные объекты и связи, в соответствии с которыми формируется сеть понятий.

Задача автоматического формирования модели предметной области еще на этапе формирования реальной системы, в процессе принятия решений об ее структуре и функциях, когда не определены объекты и связи, может рассматриваться как обратная задача, в которой сам состав понятий и связей, выбираемых при создании предметной области, определяют дальнейший состав (а иногда и функции) системы.

Именно в решении такой обратной задачи заключается проблема формирования онтологий предметных областей. Онтология может формироваться экспертами (что и происходит чаще всего), однако, перспективным кажется автоматическое формирование онтологий, которое может базироваться на знаниях, заложенных учеными, специалистами, экспертами в таких источниках, как:

- фактографические базы данных/знаний;
- текстовые корпусы;
- ресурсы современных социальных, наукометрических, библиографических сетей.

Конечно, если существуют фактографические базы данных или знаний, то проблема автоматического формирования онтологий на их основе может показаться несложной. Но при этом задача предварительного формирования таких баз данных и знаний, сама по себе, является проблемной, хотя и уже традиционной.

2. Источники формирования онтологий

Рассмотрим случаи использования двух других источников для формирования моделей предметных областей, а точнее, для частичной задачи – формирования терминологической основы предметной области систем организационного управления, связанных со свойствами живучести и надежности [Додонов, 2011]. В начале рассмотрим

построение сетей иерархий терминов на основе анализа корпуса текстов по выбранной проблематике, которое базируется на применении компактифицированных графов горизонтальной видимости для терминов – отдельных слов, биграмм и триграмм, а также установлении связей между терминами.

Как терминологическая основа для формирования терминологических онтологии используется сеть естественной иерархии терминов (СЕИТ), которая базируется на информационно-значимых элементах текста, опорных словах и словосочетаниях, методология выявления которых приведена в [Ландэ, 2014-1], [Ландэ, 2014-2]. Данная методика предусматривает реализацию шагов, охватывающих предварительную обработку исходного текста, выявление терминов, выбор из них необходимого количества наиболее весомых, непосредственное построение сети и ее отображение. На первом этапе формируется исходный текстовый корпус. Как пример такого корпуса рассматриваются полные тексты научных статей, посвященных проблематике живучести в информационных и технических системах, представленных на русском языке. В состав текстового корпуса было в качестве примера включено около 50 научных статей общим объемом около 1 млн. символов. Предварительная обработка такого текстового корпуса предусматривала выделение фрагментов текстов (отдельных статей, абзацев, предложений, слов), исключение нетекстовых символов, отсечение флективных окончаний (стемминг). На втором этапе каждому отдельному термину из текста (слову-униграмме, биграмме или триграмме) ставится в соответствие оценка их "дискриминантная сила", а именно TFIDF, которая в каноническом виде равна произведению частоты соответствующего термина (Term Frequency) в фрагменте текста на двоичный логарифм от величины, обратной к количеству фрагментов текста, в которых этот термин встретился (Inverse Document Frequency). Для последовательностей терминов и их весовых значений по TFIDF затем строились компактифицированные графы горизонтальной видимости (CHVG) и выполнялось переопределение весовых значений слов уже по этому алгоритму. Данная процедура позволяет учитывать в дальнейшем кроме терминов с большой дискриминантной силой также высокочастотные термины, которые имеют большое значение для общей тематики. На последнем этапе формирования СЕИТ осуществляется ее отображение программными средствами анализа и визуализации графов. Для загрузки СЕИТ в базы данным формируется матрица инцидентности общепринятого формата csv. На рис. 2 представлена небольшая сеть естественной иерархии терминов.

Для работы с третьим типом источников предлагается методика построения моделей предметных областей на основе зондирования информационных сетей [Ланде, 2015]. Как такая сеть рассматривается сеть понятий, соответствующих тегам сервиса Google Scholar Citations (http://scholar.google.com/citations). Множество тегов (понятий) в этой сети образуют сеть, узлы которой соответствуют понятиям, а связи – некоторую семантическую связь между ними.

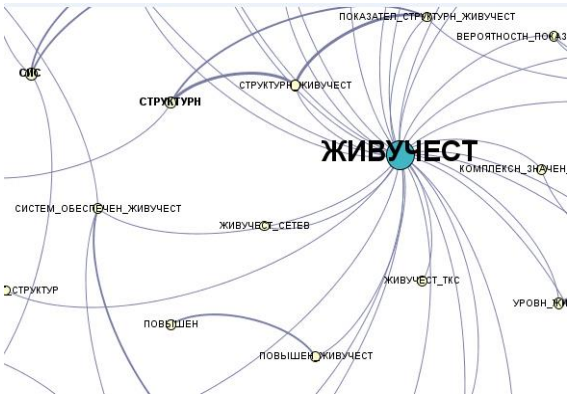

Рисунок 2 –Визуализация связного фрагмента СЕИТ

Зондирование выбранной опорной сети Google Scholar Citations осуществляется по такому алгоритму:

1. Выбирается определенное количество узлов опорной 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. Если перечень базовых узлов завершен, считается, что сеть, соответствующая результатам зондирования, построена.
5. Формирование базового стартового перечня узлов-понятий и правил отбора «конечных» узлов выполняется экспертами в предметной области.
6. Для построения модели предметной области (в рассматриваемом примере для области искусственного интеллекта) экспертным путем были определены базовые теги на английским языке: survivability, reliability, dependability, resilience, fault tolerance, availability, safety, durability, security, resiliencyи др.

На рис. 3 приведен пример архитектуры сети понятий предметной области, построенной в соответствии с приведенным алгоритмом по указанным базовым тегам.

Построенная сеть понятий оказалась связной. При количестве базовых тегов 20, общее количество узлов-тегов, которые были охвачены алгоритмом, составили 125, а количество нетерминальных узлов – 1450.

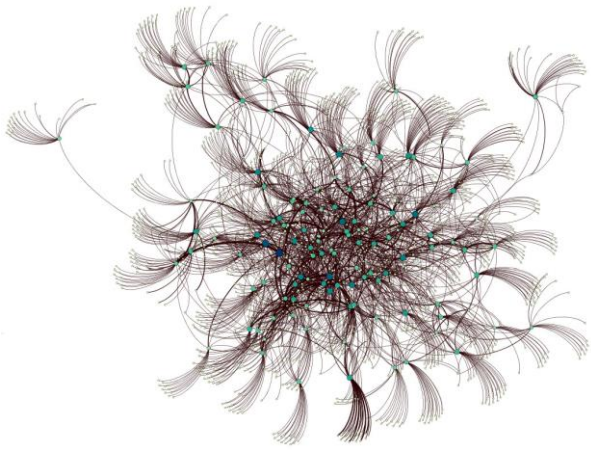

Рисунок 3 – Фрагмент архитектуры сети понятий

3. Интеграция сценарного анализа и мониторинга информационного пространства

Задача создания онтологии понятий уже рассмотрена выше, остановимся подробнее на средствах формирования сценариев и задачах мониторинга и последующей аналитики информационного пространства.

3.1. Определение возможных сценариев

Сценарии информационной поддержки, как правило, связывают с определенными факторами безопасности (чаще объектами и уязвимостями). После выбора целевых факторов сценария в графе онтологии оказываются подграфы (частичные онтологии), тесно связанные с выбранными факторами. Далее решается задача частичной оптимизации целевой функции на избранных подграфов, то есть вычисляется целевая функция в зависимости от изменений факторов безопасности, соответствующих возможным сценариям.

Существует несколько походов к вычислению взаимных влияний между узлами семантической сети [Робертс, 1986], которая может трактоваться как сеть причинно-следственных связей между объектами в заданной предметной области. В частности, в случае задачи, связанной с обеспечением безопасности некоторого объекта, соответствующего узлу, из общей каузальной сети выбираются узлы-объекты, имеющие наибольший вес связи с целевым объектом. Выделение подсетей из данных узлов и целевого узла, а также связей между ними, определение планов очередности влияния на эти подсети, задает сценарии информационной поддержки, в частности, в задачах принятия решений в области безопасности.

Детально рассмотрим подход, предложенный в [Снарский, 2015]. При решении этой задачи решающим фактором оказывается определение силы связи между отдельными узлами онтологии – между отдельными факторами безопасности, в рассматриваемом случае.

В рассматриваемой сети узлы связаны между собой связями, отражающими влияние одного узла на другой. Каждой связи приписано две характеристики: первая – направление влияния, вторая – величина влияния (рис. 4а).

Так например, узел 1 напрямую влияет на узел 3, но узел 3 напрямую на узел 1 не влияет.

Кроме того, каждой связи, задающей влияние, соответствует ее весовая величина. Например, для связи между узлами 1 и 3 это влияние задается как число ε_{13}, которое может быть как положительным, так и отрицательным.

Для решения задачи определения взаимного влияния узлов кажется естественным применение законов электротехники. Вместе с тем, подход опирающийся на непосредственное применение закона Кирхгофа приводит к некоторому логическому противоречию. В некоторых схемах влияние узла 1 на узел 2 $\varphi(1 \to 2)$ зависит от влияния узла 2 на узел 1 $\varphi(2 \to 1)$.

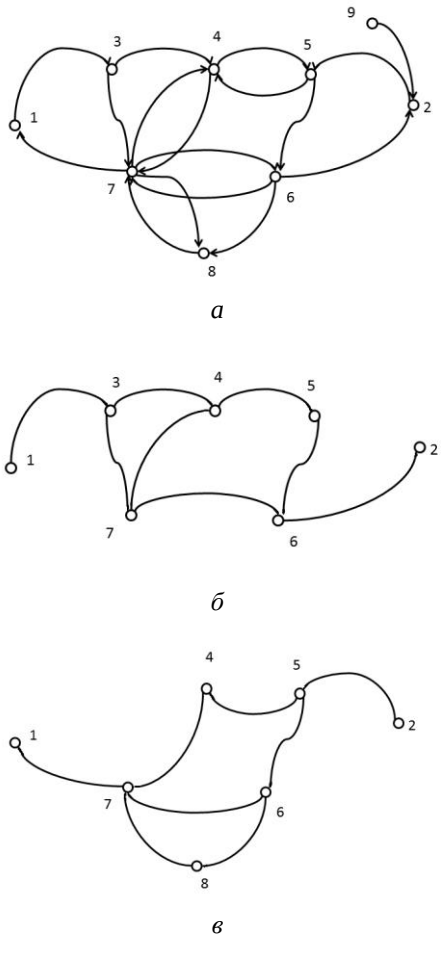

Рисунок 4 – Пример сети влияния: а – общая направленная сеть; б – подсеть влияния узла 1 на узел 2; в – подсеть влияния узла 2 на узел 1. Источник – [Снарский, 2015]

Поэтому авторами был предложен алгоритм определения величины влияния одного узла (например, 1-го) на другой (например, 2-й), состоящий из двух частей. Вначале необходимо выделить ту часть сети, по которой осуществляется

влияние первого узла на второй. Для этого из начальной сети (рис. 4а) нужно удалить все связи, оставляя только те, по которым можно пройти из первого узла во второй (рис. 4*б*). Аналогично выделяется сеть влияния второго узла на первый (рис. 4*в*). Отметим, что в общем случае сети, изображенные на рис. 4*б* и 4*в*, не совпадают.

После выделения подсетей (подсети $1 \to 2$ и $2 \to 1$) используем аналогию с электрическими сетями. Величина влияния $i-$го узла на $j-$й ε_{ij} в данной аналогии – это ЭДС. Сопротивление каждой связи в рамках данного допущения принимается равным 1. Распределение токов – J_{ij}, текущих по связям ij, определяется решением уравнения Кирхгофа:

$$\forall k: \sum_{i \to k \to j} J_{ij} = 0, \quad \forall z: \sum_{i,j \in z} J_{ij} = \sum_{i,j \in z} \varepsilon_{ij}, \qquad (1)$$

где первое уравнение требует (для стационарного случая), чтобы алгебраическая сумма токов в каждом узле k была равной нулю, а второе, чтобы для любого замкнутого контура z сумма токов (с учетом того, что сопротивление всех связей равно единице) было равно сумме всех ЭДС ε_{ij} в этом же контуре.

После решения системы уравнений (1) и нахождения токов J_{ij} выбирается любой контур, соединяющий узлы 1 и 2, например, контур 1-3-4-5-6-2 и из уравнения Кирхгофа для незамкнутого контура:

$$\sum \varepsilon + \varphi_1 - \varphi_2 = \sum J, \qquad (2)$$

находится разность потенциалов $\varphi_1 - \varphi_2$ (суммирование производится по связям выбранного контура).

Аналогично происходит вычисление для любых других пар узлов.

Именно поэтому для нахождения влияния одного узла на другой расчет приводится в соответствии с несколько измененной схемой, соответствующей рис. 4*а*–4*в*. На первом шагу от направленной сети (рис. 4*а*) происходит переход к ненаправленным подсетям (рис. 4б и 4в). Итак, на первом шаге происходит модификация направленной сети связи. После выбора узлов, для которых изучается взаимное влияние, остаются только связи, по которым можно пройти от одного узла к другому (рис. 4*б* и 4*в*).

На втором шаге происходит расчет $\varphi(1 \to 2)$ происходит по правилам Кирхгофа (рис. 4б). Естественно, при расчете влияния узла 2 на узел 1 расчет происходит по другой схеме (рис. 4*в*). При этом связи считаются обычными проводниками с единичным сопротивлением, и в $\varphi(\alpha \to \beta) = \varphi(\beta) - \varphi(\alpha)$ принимается $\varphi(\alpha) = 0$. Заметим, что при таком методе расчета влияние узла 1 на 2 не зависит от влияния узла 2 на 1.

Анализ динамики изменения значений отдельных факторов безопасности и связей во времени (как по их отражению в информационном пространстве, так и внесенных экспертами) позволяет выявлять некоторые закономерности изменения этих факторов (периодичности, тренды, аномалии) путем применения современных средств цифровой обработки сигналов (регрессионный, дисперсионный, вейвлет-анализ и т.д.), выявлять возможные информационные операции путем сравнения с соответствующими шаблонами их динамики, а также осуществлять прогнозирование [Додонов, 2014].

3.2. Определение Контент-мониторинг и аналитика

Сегодня Интернет образует значимый динамичный сегмент информационного пространства, информационные потоки, содержание и объемы которых необходимо учитывать при проведении аналитических исследований. Основным объектом анализа при этом является событийные или тематические срезы этих потоков - массивы информационных сообщений, документов, соответствующих определенным событиям или тематикам.

Задача подсистемы мониторинга информационного пространства следующие:

- мониторинг целевых объектов;
- нахождение релевантных тематических сообщений в информационном пространстве;
- контроль медиаприсутствия и медиаактивности целевых объектов;
- выявление новых объектов мониторинга;
- формирование ретроспективных фондов для последующего анализа.

Подсистема аналитической обработки представляет собой аналитический блок системы информационной поддержки, обеспечивает решение следующих задач:

- определение динамики тематических сюжетов;
- определение критических точек в динамике тематических сюжетов;
- отслеживание сюжетных цепочек, соответствующих событий, процессов;
- выявление основных событий и объектов с тематического сюжета;
- выявление и визуализация взаимосвязей событий и объектов мониторинга, а также объектов мониторинга между собой.

В соответствии со своим назначением данная подсистема, вместе с подсистемой мониторинга информационного пространства, позволяет реализовать следующие этапы информационно-аналитического исследования:

- формирование запроса в среде выбранной системы. Нахождение тематических публикаций по запросу с помощью систем контент-мониторинга;
- определение динамики тематических публикаций по запросу;
- определение критических точек в динамике тематических публикаций;
- определение основных событий в критических точках;
- выявление объектов мониторинга;
- выявление и визуализация взаимосвязей;
- прогноз развития событий.

Заключение

Таким образом, представлена архитектура системы информационной поддержки принятия решений, идеология создания и использования онтологий для построения сюжетов информационного противодействия. Также рассмотрена методика аналитического исследования, основанная на использовании инструментальных средств анализа и визуализации информационных потоков, соответствующих временных рядов и сетевых структур.

Автоматически сформированные терминологические основы онтологий и соответствующие, пусть и простейшие, семантические связи по выбранным свойствам (живучести, надежности) могут использоваться, в частности, в качестве «единого» для всех участников разработки языка предметной области, для обучения, тренингов, организации семантического поиска (организации контекстных подсказок информационно-поисковых систем), навигации пользователей в соответствующих информационных ресурсах.

Предложенные архитектурные решения можно использовать при реализации систем информационной поддержки принятия решений, основанных на контент-мониторинге информационного пространства и сценарном анализе, а так же в качестве базы для проведения аналитической и прогнозной деятельности.

Авторы благодарны коллегам и соавторам А.А. Снарскому, С.М. Брайчевскому, В.Г. Путятину, В.А. Додонову и В.Н. Фурашеву за внимание, поддержку и интерес, проявляемые при обсуждении рассматриваемых подходов.

Библиографический список

[Горбулін, 2009] Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операции та безпека суспільства: Загрози, протидія, моделювання: монографія. – К .: Інтертехнологія, 2009. – 164 с.

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

[Додонов, 2011] Додонов А.Г. Живучесть информационных систем / А.Г. Додонов, Д.В. Ландэ. – К.: Наук. думка, 2011. – 256 с.

[Додонов, 2012] Додонов А.Г., Ландэ Д.В. Методика аналитического исследования динамики событий на основе мониторинга веб-ресурсов сети Интернет // Информационные технологии и безопасность: основы обеспечения информационной безопасности: Материалы международной научной конференции ИТБ-2014. – К.: ИПРИ НАН Украины, 2014. – С. 3-17.

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

[Ландэ, 2014-1] Ландэ Д.В. Построение терминологической сети предметной области / Д.В. Ландэ, А.А. Снарский, В.Г. Путятин // Реєстрація, зберігання і обробка даних, 2014. – Т. 14. – № 2. – С. 114-121.

[Ландэ, 2014-2] Ландэ Д.В. Применение КГГВ-алгоритма для научных текстов / Д.В. Ландэ, А.А. Снарский, Е.В. Ягунова // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2014): материалы IV междунар. науч.-техн. конф. (Минск 20-22 февраля 2014 года) / – Минск: БГУИР, 2014. – С. 199-204.

[Ландэ, 2015] Ландэ Д.В. Построение модели предметной области путем зондирования сервиса Google Scholar Citations // Онтология проектирования, 2015. - N 3(17). - С. 328-335.

[Палагин, 2012] Палагин А.В. Онтологические методы и средства обработки предметных знаний: монография / А.В. Палагин, С.Л. Крывый, Н.Г. Петренко. – Луганск: изд-во ВНУ им. В. Даля, 2012. – 324 с.

[Робертс, 1986] Робертс Ф.С. Дискретные математические модели с приложениями к социальным, биологическим и экономическим задачам. - М.: Наука, 1986.

[Снарский, 2015] Снарский А.А., Ландэ Д.В. Метод выделения подсетей в каузальных сетях в задачах сценарного анализа // Информационные технологии и безопасность. Материалы XV Международной научно-практической конференции ИТБ-2015. - К.: ИПРИ НАН Украины, 2015. - С. 212-215.

[Шульц, 2011] Шульц В.Л., Кульба В.В., Шелков А.Б., Чернов И.В. Сценарный анализ эффективности управления информационной поддержкой государственной политики России в Арктике. // Национальная безопасность / nota bene. – 2011. – № 6. – С. 104-137.

[Шульц, 2014] Шульц В.Л., Кульба В.В., Шелков А.Б., Чернов И.В. Структурно-динамический подход к сценарному анализу процессов информационного противоборства в Арктике // Труды XII Всероссийского совещания по проблемам управления (ВСПУ 2014). – М.: ИПУ РАН, 2014. – С. 8889-8901.

MODELS OF SUBJECT DOMAINS IN DECISION-MAKING SUPPORT SYSTEMS ON THE BASIS OF INFORMATION SPACE MONITORING

Dodonov A.G., Lande D.V., Kovalenko T.V.

Institute for Information Recording NAS of Ukraine, Kiev , Ukraine

dodonov@ipri.kiev.ua

dwlande@gmail.com

2005ste@ukr.net

Approaches to creation of models of subject domains in decision-making support systems are described. As sources of knowledge text cases and resources of modern scientometric services are considered. The architecture of system of information support of decision-making, ideology of creation and use of ontologies for creation of plots of information support of decision-making is presented. The methodology of analytical research which is based on use of tools of the analysis and visualization of information streams and network structures is offered.

Keywords: *decision support, ontology, domain model, information space, content-monitoring*