УДК 004.891

ПОДХОДЫ К АВТОМАТИЗАЦИИ ПРОЦЕДУР ПОЛУЧЕНИЯ И ОБРАБОТКИ ЭКСПЕРТНЫХ ЗНАНИЙ НА ОСНОВЕ МОДЕЛЕЙ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Симанков Владимир Сергеевич,
д.т.н., профессор

Тарасов Елизар Савич, к.т.н.
Институт информационных технологий и безопасности Кубанского государственного технологического университета,
Краснодар, Россия

В статье рассмотрены модели и методы автоматизации процедур обработки экспертных знаний на основе алгоритмов интеллектуальной обработки данных, что позволяет повысить уровень формализации отдельных этапов экспертиз в составе ситуационных центров.

Ключевые слова: ФОРМАЛИЗАЦИЯ, ЭКСПЕРТИЗА, СИТУАЦИОННЫЙ ЦЕНТР, ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, ОБРАБОТКА ЗНАНИЙ

UDC 004.891

MATHEMATICAL MODEL FOR SYNTHESIS OF INTEGRATED SECURITY SYSTEMS

Simankov Vladimir Sergeevich,
Dr.Sc (Tech.), professor

Tarasov Elizar Savvich, Dr.Sc (Tech.)
Institute of Information Technology and security of the Kuban State Technological University
Krasnodar, Russia

The article deals with automation models and methods of expert knowledge processing procedures on the basis of intellectual data processing algorithms that allows to increase formalization level of some expert estimation procedures stages as a part of the situational centers.

Keywords: FORMALIZATION, EXPERTISE, SITUATIONAL CENTRE, INTELLECTUAL DATA ANALYSIS, DATA MINING

Подходы к автоматизации процедур получения и обработки экспертных знаний на основе моделей интеллектуального анализа данных

Актуальной задачей в области принятия решений органами государственной власти является своевременное, стабильное и эффективное информационное обеспечение всех участников и всего набора процедур принятия решений. В этой связи необходима организация и функционирование целого комплекса отдельных подсистем в рамках единой платформы для оперативного получения требуемой информации, организации эффективного взаимодействия участников процесса принятия решений и контуров обратной связи по «ключевым точкам». Описанные задачи являются чрезвычайно актуальными и своевременными, требуя для решения применения системного подхода и его реализации на базе современных информационных технологий. В такой ситуации наиболее

эффективным средством для комплексного анализа, процедур информационного обеспечения, оценок и мониторинга в рамках принятия решений выступает Ситуационный центр, построенный на платформе интеллектуальной информационно-аналитической системы.

В настоящее время системы поддержки решений и методы ситуационного управления стали развиваться в направлении адаптации к сложной динамике развития политических, экономических и социальных управленческих ситуаций. Современные системы поддержки принятия решения в большинстве случаев функционируют в условиях нечёткости и противоречивости исходной информации. В этой ситуации становятся актуальными вопросы, связанные с описанием и формализацией проблемы, эффективного подбора экспертов с учётом специфики проблемной области и представления информации заинтересованным лицам для её последующей обработки и анализа.

Существующие методы и алгоритмы для решения указанных вопросов либо отсутствуют или находятся на стадии разработки, либо недостаточно эффективны в использовании. В связи с этим особенно актуальным становится ряд проблем:

- недостаточная эффективность процессов формализации проблемы, описанной на естественном языке,
- недостаточная эффективность процедур, связанных с формализацией знаний об экспертах для последующего формирования проблемно-ориентированных экспертных групп,
- недостаточная эффективность представления, визуализации и интерпретации получаемых данных и экспертных знаний.

В этой связи нами для детального исследования были поставлены следующие цели и задачи, соответствующие указанным проблемам:

1. Повышение эффективности использования методов формализации проблемы описанной на естественном языке

- Разработка методик морфологического, синтаксического и лингвосемантического анализа описания проблемы на естественном языке.
- Разработка методики формирования набора ключевых слов (тезауруса проблемы)
- Разработка методики построения семантической сети (формальное представление проблемы)

2. Повышение эффективности процедур, связанных с формализацией знаний об экспертах для последующего формирования проблемно-ориентированных экспертных групп

- Разработка методик морфологического, синтаксического и лингвосемантического анализа анкетной информации об экспертах (сфера научных интересов, тематика публикаций, опыт проведения экспертиз и т.д.).
- Разработка методики формирования набора ключевых слов характеризующих сферу деятельности эксперта (тезауруса эксперта)
- Разработка методики построения семантической сети (формальное описание эксперта в рамках модели специалиста)

3. Повышение эффективности представления, визуализации и интерпретации получаемых данных и экспертных знаний

- Разработка методик визуализации данных, использования когнитивных графических образов и использования

динамических интерактивных сред в процессе формировании мнений экспертов и ЛПР.

- Разработка методики построения и применения когнитивных моделей в рамках совместного использования естественного и формального описаний в процедурах формирования и представления промежуточных и итоговых результатов отдельных этапов принятия решений
- Разработка методики интерпретации информации (информационных потоков), возникающих в процессе взаимодействия участников информационного обмена в ходе принятия решений.

При решении указанных задач необходимо учитывать специфику как процесса принятия решений в целом, так и отдельных процедур между участниками процесса принятия решений (и отдельными подсистемами СППР), а именно:

- Слабая (частичная) формализуемость обрабатываемой информации.
- Высокая степень разнородности информационных потоков в СППР и отсутствие (слабое развитие) интерфейсов взаимодействия между ними, а также сложности представления информации в единой форме на всех этапах принятия решений.
- Необходимость совместного использования разнородных режимов (регламентов) обработки данных;
- Недостаточная степень автоматизации процессов интеграции экспертных знаний в контур СППР

На основе анализа выявленных особенностей представляется целесообразным использование нечёткого подхода в решении

поставленных задач как платформы для применения методов и подходов лингвосемантического анализа и нечёткого когнитивного моделирования.

На этапе предварительной обработки и предметной классификации будем рассматривать экспертную информацию в ЕЯ-форме как текст, «набор слов», используя численные характеристики употребления тех или иных терминов, вне зависимости от порядка их употребления. Тогда вероятность того, что термин w, принадлежащий формируемому тезаурусу W, встречается в описании проблемы или корпусе анкет экспертов d (множества D тематического классификатор), т.е. принадлежит той или иной предметной области t:

$$P(w|d) = \sum_{t \in T} P(w|t)P(t|d) \quad (1),$$

где t – элемент множества T предметных областей.

Для оценки максимального правдоподобия параметров модели, зависящей от скрытых переменных, используем EM-алгоритм. Параметры предварительного семантического анализа $P(w|t)$ и $P(t|d)$ определим следующим образом. Пусть r – число итераций. На E-шаге вычислим $P(t|w,d)^{(r)}$:

$$P(t|w,d)^{(r)} = \frac{P(w|t)^{(r-1)} P(t|d)^{(r-1)}}{\sum_{t' \in T} P(w|t')^{(r-1)} P(t'|d)^{(r-1)}} \quad (2)$$

На M-шаге оценим параметры:

$$P(w|t)^{(r)} = \frac{\sum_{d \in D} N(w|d)\, P(t|w,d)^{(r)}}{\sum_{w \in W} \sum_{d \in D} N(w'|d)\, P(t|w',d)^{(r)}} \quad (3)$$

$$P(t|d)^{(r)} = \frac{\sum_{w \in W} N(w|d)\, P(t|w,d)^{(r)}}{\sum_{t' \in T} \sum_{w \in W} N(w|d)\, P(t'|w,d)^{(r)}} \quad (4)$$

где $N(w,d)$ – число вхождения элемента тезауруса w в рассматриваемый текст d. Описанный процесс обучения повторяется до

сходимости параметров. Однако при использовании данного алгоритма параметры часто попадают в область локального оптимума, соответственно, эффективность модели не улучшается в результате обучения. Введен дополнительный параметр $0<\beta\leq 1$ для управления скоростью обучения. Выражение для М-шага примет вид:

$$P(t|w,d)^{(r)} = \frac{\left(P(w|t)^{(r-1)}P(t|d)^{(r-1)}\right)^{\beta}}{\sum_{t'\in T}(P(w|t')^{(r-1)}P(t'|d)^{(r-1)})^{\beta}} \quad (5)$$

Для достижения глобального оптимума изначально принимаем β=1 с последующим уменьшением посредством умножения на $0<\eta<1$, пока получаемые оценки правдоподобия не улучшатся.

Определим суммарные вероятности $W(w,t)$ и $D(d,t)$ следующим образом:

$$W(w,t)^{(r)} = \sum_{d\in D} N(w,d)P(t|w,d)^{(r)} \quad (6)$$

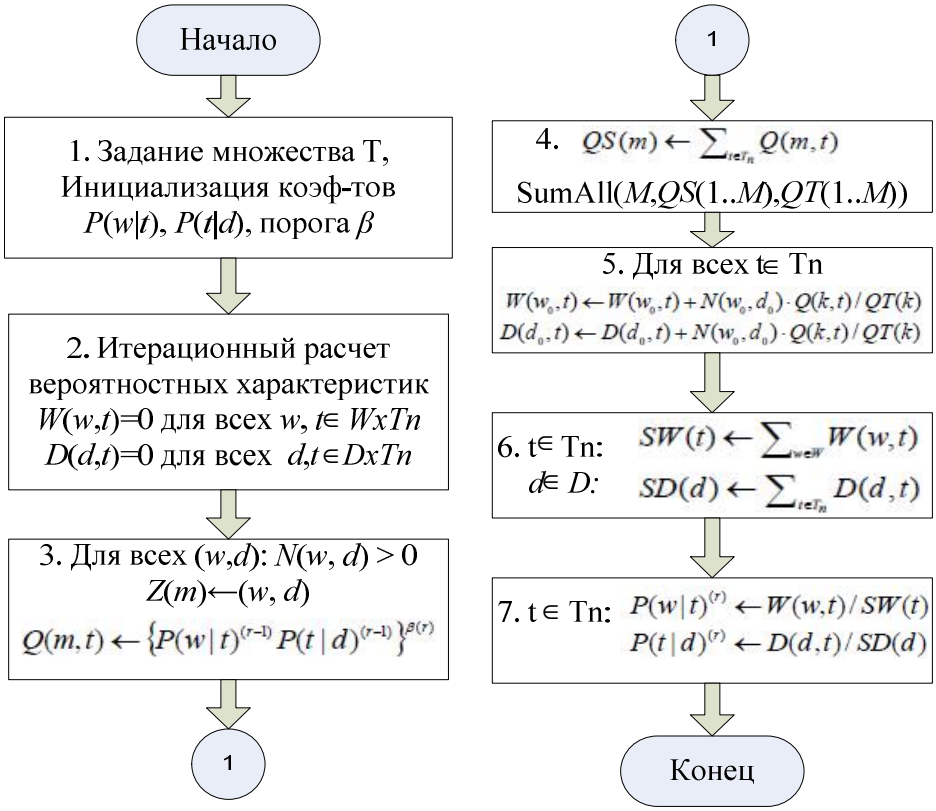
$$D(d,t)^{(r)} = \sum_{w\in W} N(w,d)P(t|w,d)^{(r)} \quad (7)$$

По формуле (5) получим:

$$W(w,t)^{(r)} = \sum_{d\in D} \frac{N(w,d)(P(w|t)^{(r-1)}P(t|d)^{(r-1)})^{\beta}}{\sum_{t'\in T}(P(w|t')^{(r-1)}P(t'|d)^{(r-1)})^{\beta}} \quad (8)$$

$$D(d,t)^{(r)} = \sum_{w\in W} \frac{N(w,d)(P(w|t)^{(r-1)}P(t|d)^{(r-1)})^{\beta}}{\sum_{t'\in T}(P(w|t')^{(r-1)}P(t'|d)^{(r-1)})^{\beta}} \quad (9)$$

Алгоритм лингвосемантического анализа примет вид (рис. 1).



T – множество предметных областей; ***M*** – число обрабатываемых (буферных) документов;
Z – массив размера M с парами ***(w, d)*** «номер термина – номер документа»;
Q(m,t) – массив для m-х промежуточных значений рассматриваемой t-области
SumAll*(m,QS,QT)* – коммуникационная процедура, получает массив *QS*, передает для вычисления суммы всех значений ото всех процессов, и возвращает их в массив *QT*.

Рис. 1 – Оптимизированный алгоритм лингвосемантического анализа с ЕМ-алгоритмом параллельного обучения

Для формирования ребер семантической сети и оценки меры семантической близости выделенных понятий (элементов тезауруса) в настоящее время используются четыре распространенных оценки: меры Jaccard, Overlap, Dice и PMI (point-wise mutual information). Эти метрики исходят из предположения, что высокие частоты совместной встречаемости терминов в тексте указывают на значительную степень ассоциации, что в свою очередь обуславливает наличие семантических связей между ними.

Для формирования итоговых обобщений имеющихся описаний и получаемой экспертной информации предложен подход, заключающийся в

формировании семантических пространств (ареалов) максимальной близости на основе применения ЕА-алгоритма к результатам лингвосемантического анализа.

Обозначим $\theta_1,\ldots,\theta_k$ – формализованная модель текста с k различными предметными областями полученной семантической сети и θ_B – модель набора текстов С. Термин w в тексте d оценивается следующей величиной:

$$p_d(w) = \lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{k}(\pi_{d,j}p(w|\theta_j)) \quad (10)$$

где w – термин в тексте d, $\pi_{d;j}$ –вес текста d для выбора j-й предметной области θ_B ($\sum_{j=1}^{k}\pi_{d,j}=1$), и λ_B – вес θ_B .

Использование модели θ_B направлено на большее разделение моделей предметных областей, т.к. θ_B присваивает высокие вероятности незначимым и неинформативным словам, снижая их влияние на модели предметных областей. θ_B оценивается на наборе текстов С и не меняется в ходе дальнейших оценок:

$$p(w|\theta_B) = \frac{\sum_{d\in C} c(w,d)}{\sum_{w\in V}\sum_{d\in C} c(w,d)} \quad (11)$$

Введем дополнительный параметр оценки $\Lambda = \{\theta_j, \pi_{d;j} | d\in C, 1\leq j\leq k\}$. Логарифмическая оценка правдоподобия C:

$$\log p(C|\Lambda) = \sum_{d\in C}\sum_{w\in V}[c(w,d)\times\log(\lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{k}\left(\pi_{d,j}p\left(w|\theta_j\right)\right))] \quad (12)$$

где $c(w; d)$ – число терминов w в тексте d.

Возникает задача найти такое значение параметра оценки Λ, которое максимизирует (12). Другими словами,

$$\Lambda = \arg\max_{\Lambda} \log p(C|\Lambda)$$

$$= \arg\max_{\Lambda} \sum_{d \in C} \sum_{u \in V} [c(w,d) \times \log(\lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} (\pi_{d,j} p(w|\theta_j))]$$

(13)

Введем «скрытые переменные», характеризующие термины: $\{z_{d,w}\}$ и $p(z_{d,w}=B)$ – вероятность того, что термин w в тексте d подчиняется выбранному фоновому распределению (модель набора текстов θ_B). $p(z_{d,w}=j)$ означает, что термин w в тексте d встречается в контексте предметной области j, и не учитывается притом общей моделью текста (не является незначимым). Получим выражения для шагов ЕМ-алгоритма.

Е-шаг:

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)}(w|\theta_{j'})} \quad (14)$$

$$p(z_{d,w} = B) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)} \quad (15)$$

М-шаг:

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)}{\sum_{j'=1}^{k} \sum_{w \in V} c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)} \quad (16)$$

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d \in C_j} c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)}{\sum_{w' \in V} \sum_{d \in C_j} c(w',d)(1-p(z_{d,w'}=B))p(z_{d,w'}=j)} \quad (17)$$

Зная оценочные параметры каждого термина, группы терминов (семантические ареалы), принадлежащих предметной области j условно будем считать "псевдотекстом", итоговым обобщением по j-й предметной области текста. Используя модель (17), мы агрегируем все семантические ареалы термина w, принадлежащего предметной области j (по всем

текстам), и нормализуем выражение $\{p(w|\theta_j)\}_{w\in V}$ для достижения $\sum_{w\in V} p(w|\theta_j)=1$.

В рамках разрабатываемой системы, как было указано выше, должны решаться следующие основные и инфраструктурные задачи:

- автоматизированные: сбор, обработка и хранение экспертных данных;
- создание и ведение БД на основании полученных экспертных знаний;
- повышение оперативности и качества управленческих решений на основе использования аналитических инструментальных средств;
- проведение мониторинга и интеллектуального анализа текущей ситуации;
- возможности визуализации информации;
- инструментальная и информационная поддержка экспертно-аналитической деятельности ЛПР и специалистов;
- обеспечение защиты, конфиденциальности и целостности информационных ресурсов системы.

На этапе практической реализации разработанных моделей, подходов и алгоритмов в рамках программного комплекса, в его составе целесообразно выделить ряд подсистем (рис. 2):

- Подсистема визуализации и представления данных (интерактивное представление данных, построение когнитивных моделей, формализация результатов, интерпретация информации);
- Подсистема формирования проблемно-ориентированных экспертных групп (подбор кандидатур с учётом специфики проблемной области на основе методик и алгоритмов анализа и формализации проблем, формализации данных об экспертах для формирования группы);

– Подсистема организации и проведения экспертиз (в том числе формирование списка вопросов к обсуждению, сбор, обработку и анализ получаемых экспертных знаний с их последующей формализацией).

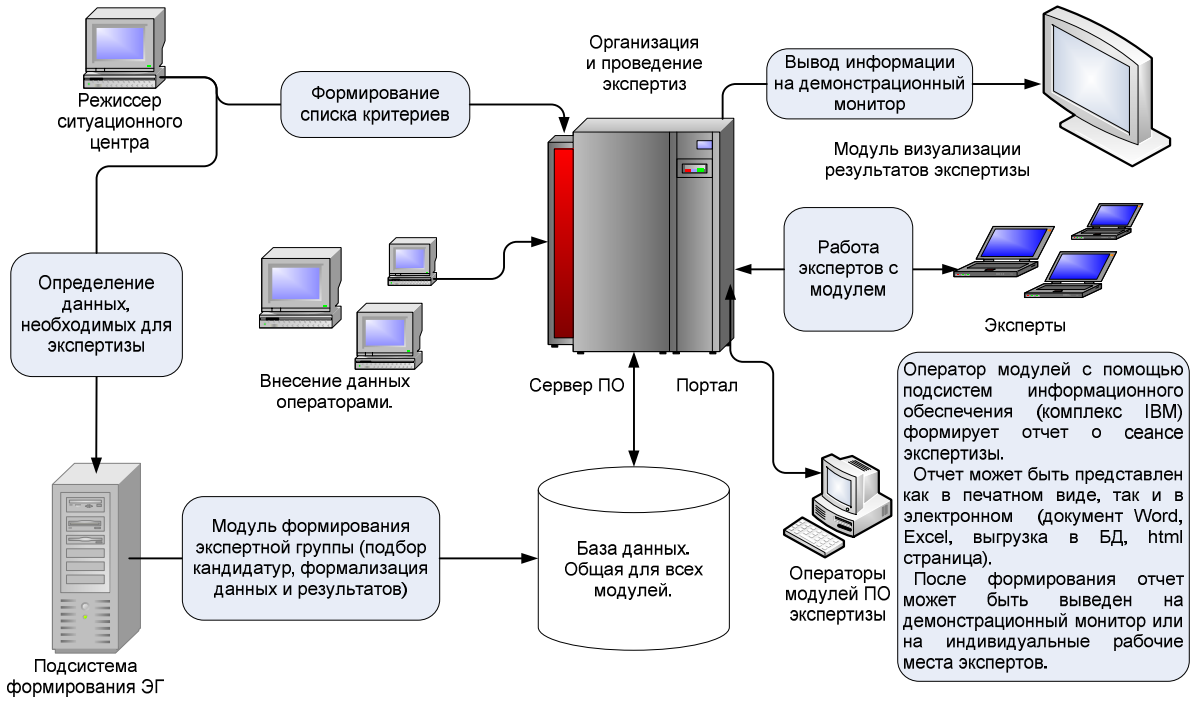


Рис. 2 – Схема взаимодействия подсистем модуля экспертизы

В итоге, по результатам анализа заключения экспертной группы возможно извлечение новых знаний с занесением их в БЗ для последующего применения в автоматизированном контуре принятия решений. Такой подход позволит проводить не только выборочную экспертизу, но и осуществлять экспертную оценку и контроль в режиме реального времени, а в случае необходимости – и постоянно: для оперативного обнаружения негативных факторов и выработки рекомендаций по их устранению с помощью сформированной проблемно-ориентированной экспертной группы. При этом вновь получаемые знаний интегрируются в базу знаний ситуационного центра, что позволяет при повторном возникновении аналогичной проблемы задействовать автоматизированный контур, что позволит экономить значительные ресурсы и время на принятие решений.

Список литературы

1. Ильин, Н.И. Новые направления развития ситуационных центров органов государственной власти/ Ситуационные центры и перспективные информационно-аналитические средства поддержки принятия решений: Матер. научно-практ. конф./ РАГС. – М.:Изд-во РАГС, 2008. – С. 12 – 16.
2. Трахтенгерц, Э.А. Субъективность в компьютерной поддержке управленческих решений. М.: СИНТЕГ, 2001. - 256 с.
3. Елагин В. В. Теоретические основы создания системы информационно-аналитического обеспечения государственного управления: диссертация доктора технических наук; 05.13.10: Челябинск, 2006. – 440 с.
4. Информационно-аналитические средства поддержки принятия решений и ситуационные центры// Материалы научно-практической конференции, РАГС, 2008 года / Под общ. ред. А.Н. Данчула. — М.: Изд-во РАГС, 2009. — 343 с.
5. Цикунов, Ю.Ф. Ситуационный центр в системе управления регионом/ Ю.Ф. Цикунов // Ситуационные центры и перспективные информационно-аналитические средства поддержки принятия решений: Матер. научно-практ. конф./ Российск. акад. гос. службы. – М.:Изд-во РАГС, 2008. – С. 16 – 20.