ОБРАБОТКА КОНТЕНТА В ИНФОРМАЦИОННЫХ СРЕДАХ НА ОСНОВЕ НЕЙРО-НЕЧЕТКОЙ МОДЕЛИ ПРИНЯТИЯ РЕШЕНИЙ

УДК 004.82

Нина Викторовна Комлева,
к.э.н., доц., проф. каф. Математического обеспечения информационных систем и инноватики, Московский государственный университет экономики, статистики и информатики (МЭСИ)
Тел.: 8 (495) 442-80-98
E-mail: nkomleva@mesi.ru

Ольга Андреевна Хлопкова,
студентка 5-го курса института Компьютерных технологий, Московский государственный университет экономики, статистики и информатики (МЭСИ)
Тел.: 8 (916) 361-60-32
E-mail: perikrone@gmail.com

Статья посвящена вопросам математического моделирования процесса принятия решения об обработке контента в информационных средах на основе нечеткой нейросети TSK. В зависимости от варьирующихся характеристик формируется интегрированная рейтинговая оценка контента, необходимая для принятия решения о его дальнейшем использовании. Для интеллектуализации поиска контента предложен механизм построения индивидуальной траектории и формирования индивидуальной компетенции.

Ключевые слова: информационная среда, интеллектуализация обработки данных, рейтинговая оценка контента, нечеткая нейронная сеть, поддержка принятия решений.

Nina V. Komleva,
PhD in Economics, Associate Professor, Professor, the Department of Software of Information Systems and Innovations, Moscow State University of Economics, Statistics and Informatics (MESI),
Тел.: 8 (495) 442-80-98
E-mail: nkomleva@mesi.ru

Olga A. Khlopkova,
Student, the Department of Computer Technologies, Moscow State University of Economics, Statistics and Informatics (MESI)
Тел.: 8 (916) 361-60-32
E-mail: perikrone@gmail.com

PROCESSING THE INFORMATION CONTENT ON THE BASIS OF FUZZY NEURAL MODEL OF DECISION MAKING

The article is devoted to the issues of mathematical modeling of the decision-making process of information content processing based on the fuzzy neural network TSK. Integral rating assessment of the content, which is necessary for taking a decision about its further usage, is made depended on varying characteristics. Mechanism for building individual trajectory and forming individual competence is provided to make the intellectual content search.

Keywords: information environment, intellectualization of data processing, assessment rating of content, fuzzy neural network, decision-making support.

1. Введение

Постоянный рост обрабатываемых объемов данных закономерен для развивающихся и информационных сред, а тенденция к увеличению информационных потоков наблюдается в подавляющем большинстве организаций. В России ежегодно увеличивается количество виртуальных библиотек, хранилищ мультимедийного контента, образовательных учреждений, имеющих доступ к сети Интернет. Возникает необходимость автоматизированной обработки контента в информационных средах любого масштаба, начиная от информационно-образовательных систем в вузах и заканчивая корпоративными средами для управления проектами. Автоматизация обработки огромных объемов данных совершенствует документооборот, рационализирует информационные потоки, и, как следствие, ведет к сокращению трудовых и стоимостных затрат [1]. Причем эффективность автоматизации целиком зависит от выбранного метода обработки контента.

Методы нечеткой логики в совокупности с концепцией нейросетей зарекомендовали себя одним из самых эффективных и перспективных инструментов для обработки контента. Применение методов из классических разделов математики, таких как линейное программирование и автоматическое управление, в ряде задач оказывается не оптимальным. Такие методы в принципе не предназначены для решения задач интеллектуальной обработки и анализа данных. В качестве их замены успешно применяются нейро-нечеткие модели. Существует обширный опыт их использования в различных предметных областях, в том числе при постановке медицинских диагнозов, обучении автопилотов, распознавании сигналов радаров, фильтрации информации, мониторинге аварийных ситуаций, анализе сейсмической активности и т.д.

2. Многокритериальная оценка контента в информационной среде

Для автоматизации обработки контента каждому материалу в информационной среде требуется сопоставить некоторый набор параметров, на основании которых принимается решение о его обработке (о перемещении между разделами, удалении). Пользователи выкладывают на web-ресурсы новые материалы, оценивают и комментируют их при помощи инструментов обсуждения [2]. Необходимость определения степени популярности и качества контента для принятия решения о его дальнейшей обработке связана с реализацией функций:

- Перенесение в хранилище контента, обладающего высокими показателями качества;
- Интеллектуальный поиск и ранжирование контента;
- Отсев и удаление не востребованного контента;
- Накопление статистики с целью модификации состава и структуры информации на портале.

Реализация вышеперечисленных опций невозможна без определения критериев оценивания материала. Особенностью информационного контента является сложность количественного оценивания его характеристик, так как универсального определения показателей качества контента не существует.

Это связано с индивидуальными особенностями каждой информационной среды и разнородностью материалов, которые в ней накапливаются, отсутствием достоверных количественно измеримых характеристик контента. Но материалы необходимо оценивать, и на основании выдвинутой оценки осуществлять их дальнейшую обработку, будь то удаление, перенесение в закрытую часть информационной среды или размещение в популярной рубрике. Очевидно, что результирующее решение строится на основании совокупности критериев, при этом некоторые из критериев зависят сразу от нескольких характеристик. Иерархия характеристик, на основании которых формируется интегрированная рейтинговая оценка, и присутствующих в наиболее распространенных типах информационных систем, выглядит следующим образом:

Суммарный рейтинг R: интегрированная оценка контента:

1. *Пользовательский рейтинг* R_u: характеризует, насколько данный материал кажется пользователям полезным, выявляется путем голосования и оценивания (*количество голосов за материал; медиана всех голосов за данный материал с учетом рейтинга каждого голоса; порог голосов, необходимый для расчета рейтинга; медиана рейтинга всех материалов на портале*).

2. *Популярность P*: показывает, как часто пользователи обращаются к материалу (*количество просмотров; обсуждаемость (количество комментариев)*).

3. *Актуальность U*.

4. *Модераторская оценка M*.

5. *Рейтинг автора контента* R_a: у каждого пользователя в системе есть свой рейтинг. Рейтинг автора может принимать дробные значения и колеблется в пределах шкалы [0, 1] (*базовый рейтинг пользователя данной категории; активность; качество обсуждений и выкладываемых материалов*).

На основании значения интегрированной оценки контента принимается решение о его обработке. Разумеется, следует избегать большого разброса значений, принимаемых параметрами, чтобы каждый из них влиял на принятие решения в зависимости от своей важности, а не порядка величины.

3. Расчет интегрированной оценки рейтинга и трудноформализуемых параметров контента

У каждого метода принятия решений есть свои достоинства и недостатки с учетом специфики задачи оценки контента: неопределенности математического описания, отсутствия формальной модели и нетривиальности. Нечеткость представления информации о контенте и формально определенных показателей его «ценности» не позволяет классическим методам математического моделирования получать адекватные количественные описания исследуемых параметров. Для обработки данных, которые являются значениями качественных характеристик без определенных шкал, некорректно применять вероятностно-статистические и детерминированные модели: на практике это приводит к неустойчивым и неадекватным действительности результатам [3].

Задача принятия решения об обработке контента эффективно решается при помощи математического аппарата нечеткой логики. Оперируя нечеткими множествами, можно делать выводы на основании параметров контента, имеющих многозначный и неточный характер:

- Трудноформализуемые критерии оценки, которые относятся к слабо структурированным сложным моделям и предполагают необходимость манипулирования нечеткими входными и выходными данными и степенью их достоверности;
- Параметры, которым сложно дать формальное определение. Например, «актуальный материал», «полезный материал»;
- Оценки, имеющие промежуточные значения: «качество материала ниже среднего», «популярность контента ближе к низкой».

Опираясь на аппарат нечеткой логики, можно оптимальным образом реализовать систему принятия решений об обработке контента в условиях неполной и нечеткой информации на основе произвольного количества параметров.

Механизм обработки контента в информационной среде можно разбить на следующие этапы:

1. Формирование значений характеристик материала, выложенного в открытой части информационной среды (общее количество просмотров, количество просмотров за неделю, оценка пользователей, рекомендации модераторов, результаты голосования на форуме).

2. При условии достижения значений характеристик заданного уровня (порог голосов, количество просмотров) рассчитывается рейтинг данного материала по формуле, учитывающей рейтинг оценивавших его пользователей, голоса модераторов и порог голосов, необходимых для рейтингового оценивания. Либо контенту присваивается фиксированный рейтинг при получении рекомендации модератора [4].

3. На основании рейтинга, рекомендаций модератора и ряда других характеристик происходит обработка контента при помощи алгоритма нечеткого вывода.

В идеальном случае пересчет пользовательского рейтинга R_u и обновление динамических рейтинговых таблиц осуществляется при каждом новом оценивании, но может происходить по расписанию или по соответствующему запросу. При этом порог количества оценок для участия в рейтинге T_i и порог исключения из рейтинга T_e фиксированы, но подвергаются корректировке в зависимости от общего количества пользователей и медианы оценок материалов.

Каждому типу пользователей (модератор, автор, зарегистрированный пользователь) соответствует вес w, складывающийся из базового значения веса для данного типа пользователя и величины k (шага изменения рейтинга) и варьирующийся в зависимости от активности пользователя:

$$w_i = w_{base} \pm k,\ w_i \in [0{,}1;\ 1],\ k = const.$$

Каждому материалу сопоставлена матрица голосов пользователей, а рейтинг материала рассчитывается эвристическим методом взвешенной суммы оценки критериев:

$$R_u = \frac{\sum_{i=1}^{k} i \sum_{j=1}^{w} c_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{w} c_{ij} + T_i} + \frac{T_i \dfrac{\sum_{l=1}^{P} R_l}{P}}{\sum_{i=1}^{k}\sum_{j=1}^{w} c_{ij} + T_i},$$

где $k = 5$ – максимальное значение оценки по шкале,
$i \in [1;5]$ – значение оценки,
$w_j \in [0{,}1;1]$ – вес пользователя, оценившего материал,
W – количество весов для данной оценки,
c_{ij} – количество оценок i с весом w_j,
R_l – рейтинг материала l,
P – количество материалов в рейтинге.

4. Предварительная обработка входных данных

Каждому материалу, хранящемуся в информационной системе, соответствует m-мерный вектор параметров, среди которых – рейтинг, оценка модератором, общее количество просмотров, количество просмотров за последнюю неделю и другие характеристики:

$$\bar{y} = \langle y_1, y_2, \ldots y_m \rangle.$$

Из векторов $\bar{y}$ формируется матрица параметров контента $Y = \|y_{ij}\|$ ($i = (\overline{1;n})$; $j = (\overline{1;m})$) размерности $n \times m$, где n – количество всех материалов, m – количество параметров. Так как дисбаланс между значениями признаков делает работу модели вывода неустойчивой, необходима корректировка матрицы.

Цель масштабирования – адаптировать диапазон значений переменной y_i для корректного расчета и достижения

однородности интерпретации входных параметров. Чтобы правильно подобрать масштаб матрицы Y, необходимо определить нижнюю и верхнюю границы отрезка, которому принадлежат значения y_i. Нижнюю границу отрезка удобнее всего принять равной нулю, то есть вычесть из всех значений матрицы значение ее минимального элемента. Верхняя граница отрезка значений определяется как среднее значение элементов матрицы. Исходные значения признаков варьируются в большом диапазоне, и работа аналитической модели с такими данными является некорректной. После процедуры нормализации все значения параметров будут приведены к одному диапазону.

Для реализации системы принятия решений о выборе контента необходимо подобрать оптимальный алгоритм нечеткого вывода. К наиболее распространенным алгоритмам нечеткого вывода на данный момент относятся алгоритмы Мамдани, Ларсена, Суджено, Цукамото [5].

При этом принимаются во внимание следующие характеристики алгоритмов:

1) Точность вывода. При прочих равных условиях погрешность алгоритма Суджено несколько меньше, чем у других алгоритмов [5, 6].

2) Скорость вывода. С вычислительной точки зрения алгоритм Суджено использует меньше ресурсов и работает на порядок быстрее, чем другие алгоритмы, так как в нем отсутствует этап дефаззификации [7].

3) Возможность представления модуля нечеткого вывода в виде нейросети. В процессе ее обучения консеквенты продукционных правил и коэффициенты функций принадлежности подвергаются корректировке. Следовательно, алгоритм нечеткого вывода должен обеспечивать подходящее представление консеквентов продукционных правил. В алгоритмах Мамдани, Ларсена и Цукамото консеквенты представлены в виде нечетких переменных с собственными функциями принадлежности термов, в алгоритме Суджено – в виде линейных зависимостей от входных параметров [6, 8].

Таким образом, алгоритм Суджено является оптимальным для задачи принятия решения об обработке контента. По алгоритму нечеткого вывода Суджено обрабатывается входной вектор значений $\bar{x}$ и формируется итоговое решение, отправлять ли материал на дальнейшую обработку.

5. Коррекция коэффициентов модуля нечеткого вывода

Использование модуля нечеткого вывода изначально не предполагает автоматической коррекции коэффициентов. Это может стать источником проблем:

1) Необходимость вмешательства эксперта в работу модуля. В сложной информационной среде с большим количеством материалов и критериев оценивания привила обработки контента постоянно меняются. Эксперт должен регулярно корректировать коэффициенты функций принадлежности лингвистических переменных и консеквенты продукционных правил. Но если представить блок нечеткого управления в виде нейросети, для корректировки параметров модулю достаточно пройти процесс обучения единожды.

2) Не исключена ситуация, в которой продукции в базе знаний изначально содержат неполные и противоречивые данные. Ошибки в коэффициентах продукционных правил сложно выявить на этапах тестирования и эксплуатации. В нейронных сетях данная проблема отсутствует, так как при обучении продукции корректируются с ориентировкой на эталонную выборку.

3) Вид и параметры функций принадлежности могут оказаться не адекватными контенту динамической информационной среды.

Применение нечеткой нейросети обосновано тем, что в отличие от регулирования блока нечеткого вывода экспертом, при котором необходимо производить периодическую перенастройку, нейронная сеть автоматизирует этот процесс. Многослойная нейронная сеть с прямым распространением сигнала – зарекомендовавшее себя средство преодоления вышеописанных ограничений систем нечеткого вывода.

6. Принятие решения об использовании контента на основании нечеткого вывода гибридной нейросети TSK

Гибридная нейронная сеть – это нейронная сеть с четкими сигналами, весами и активационной функцией, но с объединением x_i и w_i, p_i и p_j с использованием t-нормы, t-конормы или некоторых других непрерывных операций, где:

x_i – входной сигнал,

w_i – вес входного сигнала,

$p_i = x_i \times w_i$, $i = 1, ..., n$ – сила реакции на активацию входа нейрона сигналом.

Модель вывода TSK (Takagi – Sugeno – Kanga) – нечеткая система адаптивного типа, обладающая наиболее высокими показателями скорости вывода [6]. В нечеткой нейросети TSK входные переменные являются четкими, а нечеткая импликация и t-норма представлены в виде нечеткого произведения. Архитектура нейронной сети, удовлетворяющей этим свойствам, строится на основании пяти слоев.

Слой 1: фаззификация [7]. Каждый нейрон реализует функцию принадлежности соответствующему терму. Слой определяет значение гауссовой функций фаззификации $\mu(x) = e^{-\left(\frac{x-c}{\delta}\right)^{2b}}$ для каждого правила. На выходе формируется степень принадлежности переменной к терму. Параметры c и σ подлежат адаптации на этапе обучения.

Слой 2: агрегирование подусловий. Слой определяет результирующую степень принадлежности α_i посредством перемножений частных весов $\mu_A^k(x_i)$. Количество входных связей равно количеству подусловий в антецеденте правила, количество нейронов – количеству продукций.

Слой 3: генератор функции консеквента каждого правила $z_i = \sum_{j=1}^{N} k_{ij} x_j + k_{i0}$. Причем функция умножается на вес α_i, значение которого определяется в предыдущем слое. Количество нейронов в 3-ем слое равно количеству продукций M. На выходе слоя формируются произведения $\alpha_i \cdot z_i$. Слой является параметрическим. В данном слое на этапе обучения корректируются коэффициенты k_j^i функций вывода z_i.

Слой 4: регрессионный слой, формирующий выходной сигнал. Содержит два нейрона-сумматора. На вход 1-го нейрона поступают значения функций консеквентов, умноженные на веса правил $\alpha_i \cdot z_i$. Нейрон вычисляет взвешенную сумму сигналов z_i. На вход второго нейрона поступают веса правил α_i.

Слой 5: слой состоит из единственного нейрона, агрегирующего четкий выходной сигнал:

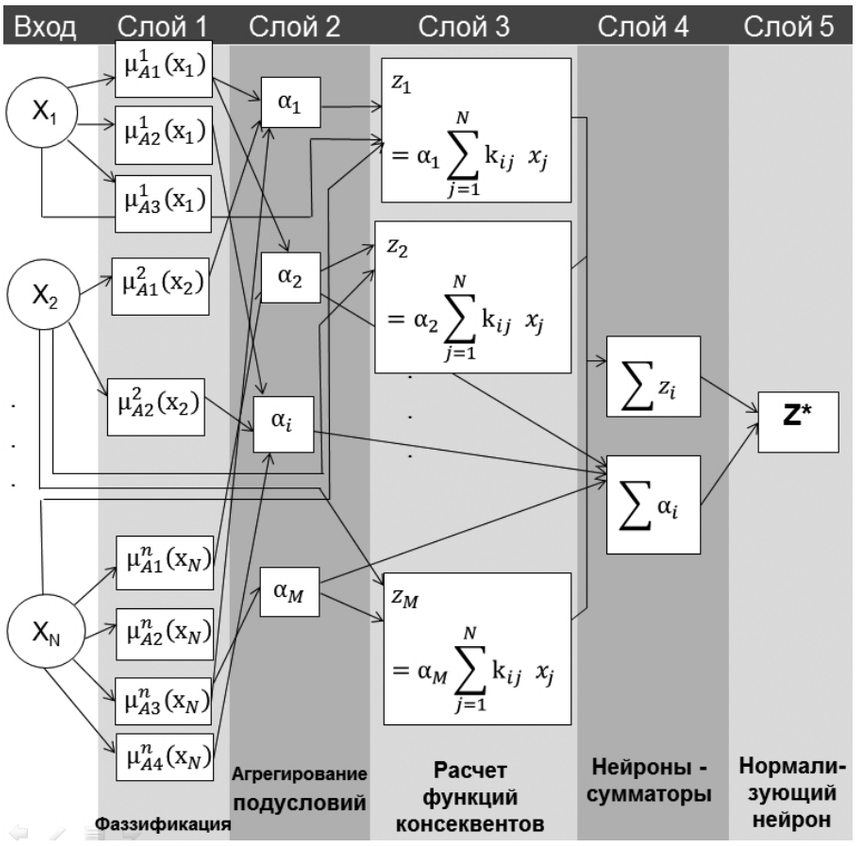

Рис. 1. Топология нечеткой сети TSK

$$z^* = \frac{\sum_{i=1}^{M} \alpha_i z_i}{\sum_{i=1}^{M} \alpha_i}.$$

Структура нечеткой нейросети TSK представлена на рисунке 1.

7. Гибридный алгоритм обучения нечеткой нейросети

Обучение сети заключается в таком подборе параметров, при котором каждый нейрон параметрического слоя будет корректно выполнять ожидаемый от него этап вывода. Алгоритм обратного распространения ошибки используется как один из этапов гибридного метода обучения для настройки параметров нейросети TSK. В процессе обучения происходит изменение параметров модуля нечеткого управления, минимизирующее меру погрешности.

Корректировке подлежат линейные коэффициенты 3-го слоя и нелинейные коэффициенты функций принадлежности 1-го слоя. Гибридный метод реализуется чередованием двух этапов, на каждом из которых корректируются параметры только одного слоя. Сначала при фиксированных значениях коэффициентов функций принадлежности рассчитываются параметры полиномиального слоя. Затем методом обратного распространения ошибки вычисляются коэффициенты c и σ 1-го слоя. После модификации линейных и нелинейных коэффициентов процесс адаптации параметров повторяется вплоть до их стабилизации. Гибридный алгоритм считается наиболее эффективным способом обучения нейросетей TSK [7].

8. Построение индивидуальных траекторий при использовании контента

Вместе с контентом в информационной среде может содержаться метаинформация, определяющая ключевые слова по содержанию и дополнительные данные. Одна из целей расчета интегрированной рейтинговой оценки и использования метаинформации – это повышение качества поиска, выборки и достижения цели пользователя при выборе контента. Помимо непосредственно поиска по заданным критериям пользователь может обратиться к модулю поддержки принятия решений как к вспомогательному инструменту экспертного анализа [9].

Для реализации данного функционала может быть выбран один из методов поддержки принятия решений. В последнее время наиболее популярным в информационно-аналитических системах является метод анализа иерархий (МАИ) [10], который основывается на декомпозиции и агрегировании: осуществляется разбиение рассматриваемых объектов на кластеры, которые разбиваются на меньшие кластеры. Целью является получение приоритетов всех элементов посредством группирования. Пусть имеются: список однотипных альтернатив – материалы в информационной среде, главный критерий сравнения альтернатив – выбор материалов, которые наиболее полно удовлетворяют запросам пользователя, и несколько групп однотипных факторов, влияющих на отбор альтернатив – параметры контента. Тогда ранжирование материалов производится на основании некоторых параметров, например, совпадение по типу и по содержанию.

Основным достоинством МАИ является возможность точной оценки и проверки ее правильности.

9. Заключение

Модуль принятия решения оперирует трудноформализуемыми понятиями («рейтинг материала ниже среднего», «очень высокая популярность»), которые должны быть определенным образом обработаны. Оперативная и корректная обработка нечетких характеристик достигается путем использовании гибридных нейросетей. Они получили широкое распространение в системах с модулями нечеткой логики.

Популярность нечетких нейросетей связана со способностью к адаптации и обучению [11, 12]. Эти достоинства нейро-нечеткого вывода в совокупности с возможностью корректировки продукционных правил на порядок повышают интеллектуальность модуля вывода. Общая схема обработки контента представлена на рисунке 2.

Таким образом, на основании значений параметров контента рассчитывается интегрированная оценка для принятия решения об обработке материалов и производится ранжирование итогов поисковых запросов пользователей.

Многокритериальный расчет интегрированной рейтинговой оценки необходим для оптимизации работы

Рис. 2. Схема обработки контента

информационной среды. Его введение упрощает ряд задач, в том числе: обработку не востребованного контента, перенесение в хранилище контента, обладающего необходимыми характеристиками, качественный поиск и ранжирование материалов, накопление статистики для модификации состава информации в информационной среде. Предложенный алгоритм интеллектуальной обработки данных может быть использован в любом проекте, где требуется рейтинговое оценивание, автоматизация процесса принятия решений, интеллектуальный поиск и ранжирование контента.

Литература

1. Могилев, А.В., Листрова, Л.В. Технологии поиска и хранения информации. Технологии автоматизации управления. – СПб: БХВ-Петербург, 2012, 320 с.

2. Комлева, Н.В. Профессиональные сообщества в системе управления знаниями // Открытое образование. Научно-практический журнал. – 2010. – №1. – С. 96–102.

3. Горяшко А.П., Комлева Н.В. Web-портал сообщества профессионалов в сфере налогообложения как инструмент управления знаниями и принятия коллективных решений // Экономика. Налоги. Право. Научное периодическое издание. – 2011. – № 4. – С. 14–21.

4. Горяшко А.П., Комлева Н.В. Образование на основе компетенций в открытых информационных средах: алгоритмы принятия решений // Высшее образование в России. Научно-педагогический журнал Министерства образования и науки РФ. – 2011. – № 8–9. – С. 78–84.

5. Борисов, В.В., Круглов, В.В., Федулов, А.С. Нечеткие модели и сети. – М: Горячая Линия – Телеком, 2012. – 284 с.

6. Круглов, В.В., Дли, М.И., Голунов, Р.Ю. Нечеткая логика и искусственные нейронные сети. – М: Физматлит, 2001. – 224 с.

7. Рутковская, Д., Пилиньский, М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. – М: Горячая Линия – Телеком, 2008. – 452 с.

8. Емельянов, С. Искусственный интеллект и принятие решений. Выпуск 1. – М: ИСА РАН, 2013. – 88 с.

9. Andreichicov, A.V., Andreichicova, O.N. New Paradigms of Decision-Making. – Halifax, Canada: Management Science Division, 2003. – pp. 28–34.

10. Саати, Т.Л. Принятие решений при зависимостях и обратных связях. Аналитические сети. – М.: Издательство ЛКИ, 2008. – 360 с.

11. Осовский С. Нейронные сети для обработки информации. – М: Финансы и статистика, 2004. – 344 с.

12. Simon S. Haykin. Neural networks and learning machines. – Hamilton, Ontario, Canada: Pearson Prentice Hall, 2009. – 906 с.

References

1. Mogilev, A.V., Listrova, L.V. Search Technology and Information Storage. Automation Technologies of Management. – St. Petersburg: BHV-Petersburg, 2012, 320 pp.

2. Komleva, N.V. Professional Communities in the System of Knowledge Management // The Open Education. Scientific and Practical Journal. – 2010. – №1. – pp. 96–102.

3. Goryashko, A.P., Komleva, N.V. Web-portal of the Professional Community in the Field of Taxation as a Tool for Knowledge Management and Collective Decision-Making // Economy. Taxes. Law. Scientific Periodical. – 2011. – № 4. – pp. 14–21.

4. Goryashko, A.P., Komleva, N.V. Education Based on Competences in Open Environments: Algorithms of Decision-Making // Higher Education in Russia. Scientific and pedagogical journal of the Ministry of education and science of the Russian Federation. – 2011. – № 8–9. – pp. 78–84.

5. Borisov, V., Kruglov V.V., Fedulov, A.S. Fuzzy Models and Networks. – M.: Hot Line – Telecom, 2012. – 284 pp.

6. Kruglov V.V., Long, M.I., Golunov, R.Y. Fuzzy Logic and Artificial Neural Networks. – M.: Fizmatlit, 2001. – 224 pp.

7. Rutkowskaya, D., Pilinski, M, L. Rutkowski. Neural Networks, Genetic Algorithms and Fuzzy Systems. – M.: Hot Line – Telecom, 2008. – 452 pp.

8. Emel'yanov, S. Artificial Intelligence and Decision-Making. Issue 1. – M: ISA RAS, 2013. – 88 pp.

9. Andreichicov, A.V., Andreichicova, O.N. New Paradigms of Decision-Making. Halifax, Canada: Management Science Division, 2003. – pp. 28–34.

10. Sahati, T.L. Decision-Making Dependencies and Feedbacks. Analytical Network. – M: Publishing house of LCG, 2008. – 360 pp.

11. Osovski S. Neural Networks for Information Processing. – M: Finance and Statistics, 2004. – 344 pp.

12. Simon S. Haykin. Neural Networks and Learning Machines. – Hamilton, Ontario, Canada: Pearson Prentice Hall, 2009. – 906 pp.