Инженерный вестник Дона, №3 (2015)
ivdon.ru/ru/magazine/archive/n3y2015/3172

Об одном способе построения запросов к базе данных на основе аппарата нечеткой логики

Н.Н. Венцов, В.В. Долгов, Л.А. Подколзина

Донской государственный технический университет, Ростов-на-Дону

Аннотация: Показано что актуальным направлением развития электронного документооборота является переход к информационным системам, поддерживающим диалоговое взаимодействие с пользователем на естественном языке. Данная проблема является актуальной по причине лавинообразного роста данных обрабатываемых современными информационными системами. На примере данных отделов кадров предприятий описана проблема преобразования словесных критериев, описывающих свойства искомых данных, в SQL-запросы, выполнение которых позволит получить требуемые информационные массивы. Математически формализация словесных критериев фильтрации данных была формализована при помощи трапецеидальных функций принадлежности нечеткому множеству. Описана реализация процесса преобразования трапецеидальной функции принадлежности в условные выражения SQL-запросов к реляционных систем управления базами данных. Реализованные преобразования математически соответствуют как теории нечетких множеств, так и концепции реляционных баз данных (т.е. стандарту SQL). Соответствие преобразований теории нечетких множеств в дальнейшем позволит использовать операторы работы с множествами (объединение, пересечение и т.д.), а соответствие требованиям SQL практически реализовывать предложенный подход средствами широкого класса реляционных систем управления базами данных.

Ключевые слова: нечеткая логика, функция принадлежности, нечеткое множество, терм, лингвистическая переменная, база данных, нечеткие запросы, α-срез, индекс соответствия

Введение

Повышение объемов производимой и перерабатываемой информации стимулирует к переходу на электронные формы организации документооборота. Использование компьютерных технологий позволяет не ограничивать обработку документов операциями чтения и записи, а использовать интеллектуальные алгоритмы формирования документации на основе имеющейся первоначальной информации. Одной из проблем, возникающей на данном этапе, является формализация качественных высказываний, оформленных на естественном языке. Работая с первичными данными пользователю удобнее оперировать инструментами родного языка, а не использовать специфический математический аппарат. Для автоматизации связи качественных харак-

теристик, заданных на естественном языке, с количественными параметрами объектов или процессов часто используют теорию нечетких множеств и нечеткую логику [1].

Актуальной задачей исследования является повышение эффективности управления сложными объектами через разработку алгоритма интеллектуализации системы, способной в той или иной степени воспроизводить действия человека, связанные с анализом, классификацией знаний в предметной области, накопленными оператором или самой системой [2 – 4].

Постановка задачи

Создание, модификация и передача электронных документов может быть реализовано только средствами автоматизированной системы (АС), где данный электронный документ приобретает статус, регистрируется и хранится. Регистрация электронных копий должна обеспечиваться средствами АС [5].

В процессе создания электронного документа средствами АС, необходимо формализовывать не только четкие количественные, но и расплывчатые качественные характеристики объектов и процессов реального мира. Рассмотрим практический пример. На некотором предприятии было решено провести сортировку сотрудников по стажу их работы. Было выделено 3 группы сотрудников: сотрудники с маленьким стажем, сотрудники со средним стажем и сотрудники с большим стажем работы, а также обратные этим высказываниям запросы. Требуется сформировать и выполнить SQL-запрос к базе данных, позволяющий выбрать сотрудников согласно указанным выше нечетким критериям.

Перед созданием базы данных введем некоторые определения: стаж сотрудников учитывается от 0-25 лет. Временной промежуток до 7 лет является показателем сотрудников с малым стажем. Максимально

приближенными к группе сотрудников со средним стажем являются те, у которых стаж работы от 8 до 13 лет; большим стажем – более 15 лет.

Предлагаемый подход

Для формализации стажа работы используем понятие лингвистической переменной. Мы определяем функцию совместимости не на множестве математически точно определенных объектов, а на множестве обозначенных некими символами впечатлений на естественном языке [6]. Лингвистическая переменная определяется набором $(X, T(X), U, G, M)$. где X стаж работы; T(X) – терм-множество X: T(Стаж работы) = «маленький» + «средний» + «большой» + «не малый» + «не средний» + «не большой»; где любое из этих значений является названием нечеткой переменной. Каждый терм накладывает некоторые нечеткие ограничения на значения базовой переменной «Стаж работы» в универсальном множестве $U \in [0;25]$. При этом, нечеткое ограничение на значения «Стаж работы» характеризуется функцией совместимости, ставящей базовой переменной число из интервала от 0 до 1, соответствующее совместимости этого значения с заданным нечетким ограничением. G – синтаксическое правило, определяющее способ порождения бесконтекстной грамматикой лингвистических значений, принадлежащих терм-множеству переменной «Стаж работы». M – семантическое правило, задающее способ вычисления смысла любой лингвистической переменной. Лингвистическое значение включает в себя первичные термы («малый», «большой» и т.д.) и связки («и», «или», «не», «очень», «более» и т.д.). Связки трактуются как видоизменяющие операторы. Так, смысл лингвистического значения «не малый» получим, вычитая из 1 значение функции совместимости терма «малый».

Любой нечеткой переменной соответствует некоторое ограничение C, представляющее собой подмножество универсального множества.

Характеристикой такого нечеткого подмножества универсального множества X выступает функция принадлежности $MF_c : X \to [0,1]$, ставящая в соответствие каждому элементу $x \in X$ число $MFc(x)$ из интервала от 0 до 1. $MFc(x)$ – характеризует степень принадлежности элемента к нечеткому подмножеству C.

Существует множество типовых форм кривых для задания функции принадлежности. В этом примере применена типовая кусочно-линейная трапецеидальная функция принадлежности [7 - 9], использующая четверку чисел *(a, b, c, d)*:

$$MF_c(x) = \begin{cases} 1 - \dfrac{b-x}{b-a}, \_ a \le x \le b \\ 1, b \le x \le c \\ 1 - \dfrac{x-c}{d-c}, c \le x \le d \\ 0, x \notin (a,d) \end{cases} \quad (1)$$

Универсальное множество X представляет собой интервал от 0 до 25. Переменная *x* принимает значения данного интервала и интерпретируется как стаж работы. Каждая функция принадлежности описывается четверкой чисел: Малый стаж – (0, 0, 4, 7); Средний стаж – (5, 8, 13, 15); Сотрудник с большим стажем – (11, 16, 25, 25).

Функция принадлежности к «Малый стаж»:

$$MF_{малый}(x) = \begin{cases} 1, \ \ 0 \le x \le 4 \\ 1 - \dfrac{x-5}{7-5}, 5 \le x \le 7 \\ 0, x > 7 \end{cases} \quad (2)$$

Функция принадлежности к «Средний стаж»:

$$MF_{средний}(x)=\begin{cases}0,\ x<5, x>15\\ 1-\dfrac{7-x}{7-5},\ 5\le x\le 7\\ 1,\ 8\le x\le 13\\ 1-\dfrac{x-14}{15-14},\ 14\le x\le 15\end{cases} \quad (3)$$

Функция принадлежности к переменной «Большой стаж» примет вид:

$$MF_{большой}(x)=\begin{cases}0,\ x<11\\ 1-\dfrac{15-x}{15-11},\ 11\le x\le 15\\ 1,\ 16\le x\le 25\end{cases} \quad (4)$$

Для удобства будем использовать срезы – фильтры по измерениям, в которых фигурируют нечеткие величины [10]. Для получения значений, принадлежащих α-срезу для каждого из терма, зададим минимальный индекс соответствия α=0,75. На основе этих данных построим график, изображенный на рисунке 1 и демонстрирующий изменение этих групп.

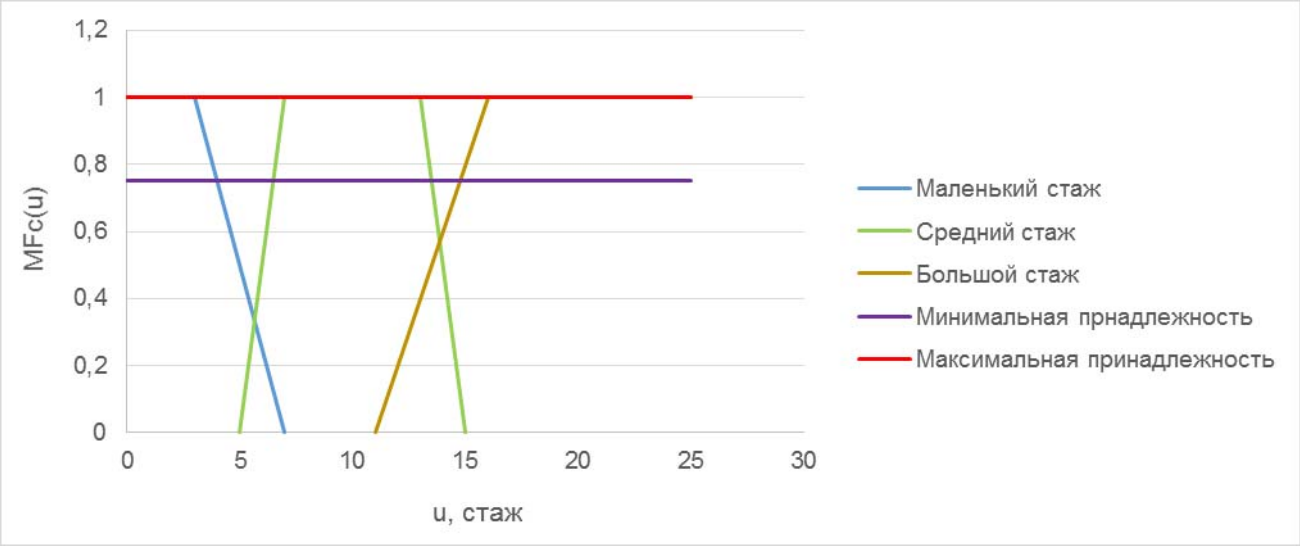

Рис. 1 – Совокупность функций принадлежности

Аналогично построим график отрицания, представленный на рис.2, для сотрудников со стажем: не маленьким, не средним, не большим. Для этого воспользуемся следующей формулой:

$$y=1-MF_c(x) \quad (5)$$

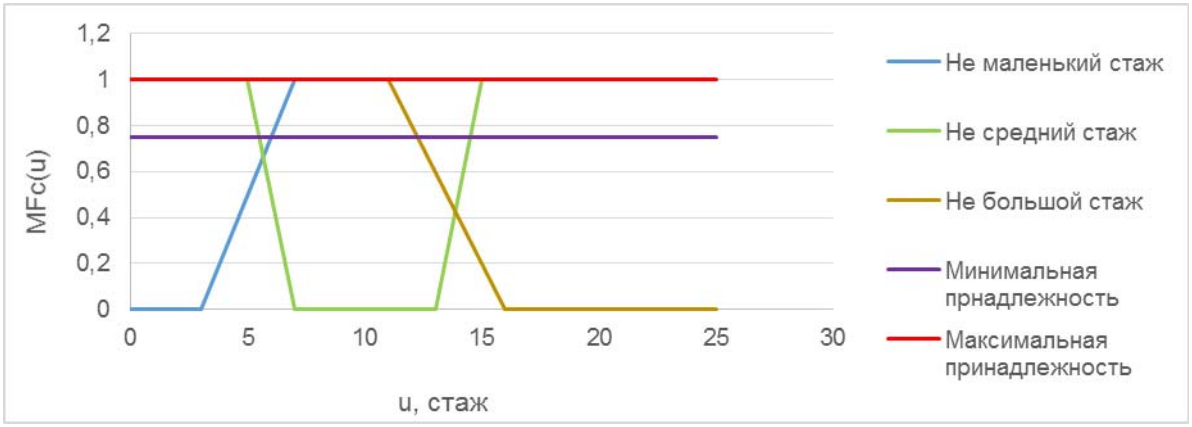

Рис. 2 – График отрицания

Следующим этапом требуется перевести описанный аппарат нечеткой логики в строгие запросы реляционных БД.

Для удобства работы рассмотрим выполнение нечеткого запроса к базе данных на основе таблицы «Сотрудники». Она состоит из следующих полей: id (уникальный идентификатор), Фамилия, Имя, Отчество, Пол сотрудника, Телефон, Примечания, Дата приема на работу, Дата увольнения, Стаж работы в аналогичной должности, Стаж всего.

Для того, чтобы узнать стаж сотрудника в данный момент времени, необходимо воспользоваться выражением, приведенным на рис.3:

```
DateDiff("yyyy",[Дата приема],Now())+[Стаж работы в
подобной должности];
```

Рис. 3 – Текущий стаж сотрудника

где Now() – функция, возвращающая текущую дату; DateDiff – возвращает разницу между двумя датами, а ее параметр «yyyy» указывает на то, что разница возвращается в годах; [Стаж работы в подобной должности] – выполняет учет стажа сотрудника до принятия его на текущую должность.

Для выбора сотрудников с маленьким стажем работы, выполним SQL-запрос к БД, приведенный на рис.4, составленный с учетом формулы (2).

```
SELECT (DateDiff("yyyy",[Дата приема],Now())+[Стаж
работы в подобной должности]) AS [Общий Стаж], *
FROM Сотрудники
WHERE ((((DateDiff("yyyy",[Дата при-
ема],Now())+[Стаж работы в подобной должности)-x)/7-
5)>=0.75) AND ((Сотрудники.[Дата увольнения]) Is
Null));
```

Рис. 4 – Выбор сотрудников с малым стажем работы

Для всех запросов верно: индекс соответствия «α=0,75» позволяет исключить результаты, лежащие ниже этого значения; «7-5» – результат ограничений, наложенных 2 формулой.

Выполнив следующий запрос из рис.5, получим данные о сотрудниках с большим стажем работы, работающих на предприятии.

```
SELECT (DateDiff("yyyy",[Дата приема],Now())+[Стаж
работы в подобной должности]) AS Стаж, *
FROM Сотрудники
WHERE (DateDiff("yyyy",[Дата приема],Now())+[Стаж
работы в подобной должности])/4>=0.75 and [Дата
увольнения] is null;
```

Рис. 5 – Запрос на выбор сотрудников с большим стажем

Аналогичным образом можно получить и остальные группы сотрудников по условию стажа работы. В общем виде алгоритм поиска сотрудников с определенным стажем работы представлен на рис. 6.

Первый шаг «Выбор вида функции принадлежности» требует: задать имя функции принадлежности, дать короткое описание (пояснения. Используемые переменные и пр.), указать тип (трапецеидальный, гауссов и т.д.), заполнить необходимые коэффициенты, для дальнейшей работы. Функция описывает стаж работы сотрудника в некоторой сфере профессиональной деятельности. В работе используется трапецеидальная функция принадлежности. Основное требование: значение функции

принадлежности должно быть больше 0 хотя бы для одного лингвистического терма.

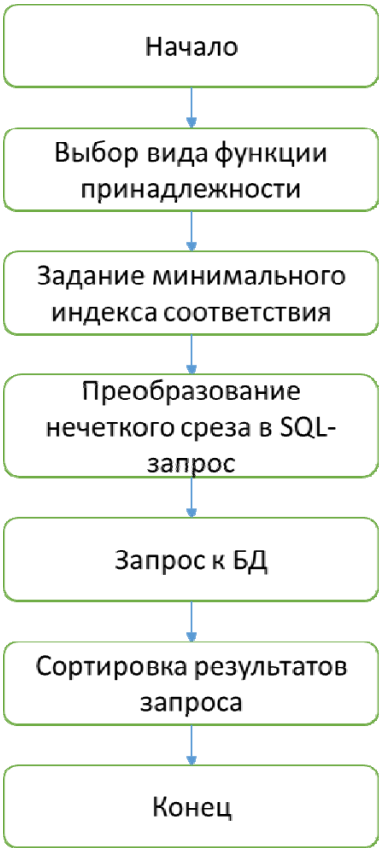

Рис. 6 – Получение результатов нечеткого запроса

На следующем шаге «Задание минимального индекса соответствия» следует выбрать минимальный индекс соответствия, обеспечивающие степень принадлежности не ниже этой границы (в данном случае значение α=0,75).

Третий шаг «Преобразование нечеткого среза в SQL-запрос»: на основе проведённых исследований предметной области, составляется запрос (правило), представляющий собой инструкцию SELECT языка SQL с обязательным указанием сотрудников, попадающих под данное правило. Критерии, согласно которым происходит отбор, задаются в установки предикатов WHERE, условие которых может быть верным или неверным для любой записи таблицы. Сложные запросы могут оформляться хранимыми

процедурами и функциями. Поэтому ограничения накладываются только возможностями SQL-сервера, на котором функционирует база данных.

Шаг «Запрос к БД» позволяет получить результат работы нечеткого запроса. Следующий шаг «Сортировка результатов запроса» необязателен, служит для распределения конечной выборки по степени соответствия заданному запросу.

Заключение

Показано что актуальным направлением развития электронного документооборота является переход к информационным системам, поддерживающим диалоговое взаимодействие с пользователем на естественном языке. Описана проблема преобразования словесных критериев, описывающих свойства искомых данных, в SQL-запросы, выполнение которых позволит получить требуемые информационные массивы. Математически формализация словесных критериев фильтрации данных была формализована при помощи трапецеидальных функций принадлежности. Описана реализация процесса преобразования трапецеидальной функции принадлежности в SQL-запросы к реляционной базе данных. Реализованные преобразования математически соответствуют как теории нечетких множеств, так и концепции реляционных баз данных (т.е. стандарту SQL). Соответствие преобразований теории нечетких множеств в дальнейшем позволит использовать операторы работы с множествами (объединение, пересечение и т.д.), а соответствие требованиям SQL практически реализовывать предложенный подход средствами широкого класса систем управления базами данных.

Нечеткий поиск в базе данных приносит аналитику максимальную пользу в случаях, когда требуется не только извлечь информацию, оперируя нечеткими понятиями, но и проранжировать ее по степени релевантности запроса [11].

Благодарность

Работа выполнена при финансовой поддержке РФФИ - проекты № 13-01-00343 и 15-01-05129.

Литература

1. Алиев Р.А., Церковный А.Э., Мамедов Г.А. Управление производством при нечеткой исходной информации. М.: Энергоатомиздат, 1991. 240 с.

2. Пивкин В.Я., Бакулин В.П., Кореньков Д.И. Нечеткие множества в системах управления. Новосибирск: НГУ, 1998. 75 с.

3. Гинис Л.А. Развитие инструментария когнитивного моделирования для исследования сложных систем // Инженерный вестник Дона, 2013, №3 URL: ivdon.ru/ru/magazine/archive/n3y2013/1806

4. Венцов Н.Н. Разработка алгоритма управления процессом адаптации нечетких проектных метаданных // Инженерный вестник Дона, 2012, №1 URL: http://ivdon.ru/ru/magazine/archive/n1y2012/630

5. Алтунин А.Е., Семухин М.В. Модели и алгоритмы принятия решений в нечетких условиях. Монография. М.: Издательство Тюменского государственного университета, 2000. 352 с.

6. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир, 1976. 166 с.

7. Бочарников В.П. Fuzzy-технология: математические основы, практика моделирования в экономике. СПб.: Наука РАН, 2001. 328 с.

8. Valiant Leslie. Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. New York: Basic Books, 2013. 208 p.

9. Novák V. “Are fuzzy sets a reasonable tool for modeling vague phenomena?” // Fuzzy Sets and Systems 156. 2005. pp. 341–348.

10. Earl Cox. Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration. Amsterdam: Elsevier/Morgan Kaufmann. 2005. 530 p.

11. Досмухамедов Б.Р. Моделирование и подходы к управлению бизнес-процессами в микрофинансовых организациях // Вестник АГТУ: Управление, вычислительная техника и информатика, 2013 №2. С.121-130.

References

1. Aliev R.A., Cerkovnyj A.Je., Mamedov G.A. Upravlenie proizvodstvom pri nechetkoj ishodnoj informacii [Production management in the fuzzy initial information]. M.: Jenergoatomizdat, 1991. 240 p.

2. Pivkin V.Ja., Bakulin V.P., Koren'kov D.I. Nechetkie mnozhestva v sistemah upravlenija [Fuzzy sets in management systems]. Novosibirsk: NGU, 1998. 75p.

3. Ginis L.A. Inženernyj vestnik Dona (Rus). 2013. №3 URL: ivdon.ru/ru/magazine/archive/n3y2013/1806

4. Vencov N.N. Inženernyj vestnik Dona (Rus), 2012, №1 URL: http://ivdon.ru/ru/magazine/archive/n1y2012/630

5. Altunin A.E., Semuhin M.V. Modeli i algoritmy prinjatija reshenij v nechetkih uslovijah [Models and algorithms for decision making in fuzzy conditions]. Monografija. M.: Izdatel'stvo Tjumenskogo gosudarstvennogo universiteta, 2000. 352 p.

6. Zade L. Ponjatie lingvisticheskoj peremennoj i ego primenenie k prinjatiju priblizhennyh reshenij [The concept of linguistic variable and its application to decision-making close]. M.: Mir, 1976. 166 p.

7. Bocharnikov V.P. Fuzzy-tehnologija: matematicheskie osnovy, praktika modelirovanija v jekonomike [Fuzzy-technology: the mathematical foundations of the practice of modeling in economics]. SPb.: Nauka RAN, 2001. 328 p.

8. Valiant Leslie. Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. New York: Basic Books, 2013. 208 p.

9. Novák V. “Are fuzzy sets a reasonable tool for modeling vague phenomena?” // Fuzzy Sets and Systems 156. 2005. pp. 341–348.

10. Earl Cox. Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration. Amsterdam: Elsevier/Morgan Kaufmann. 2005. 530 p.

11. Dosmuhamedov B.R. Vestnik AGTU: Upravlenie, vychislitel'naja tehnika i informatika, 2013 №2. pp.121-130.