УДК 004.55

Д.С. Негурица, Т.Б. Шатовская

Харьковский национальный университет радиоэлектроники, Харьков

КЛАССИФИКАЦИЯ ПОЛЬЗОВАТЕЛЕЙ В АДАПТИВНЫХ ИНТЕРФЕЙСАХ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ WEB-ОРИЕНТИРОВАННЫХ СИСТЕМ

В статье выполнен анализ известных методов динамической классификации пользователей в адаптивных интерфейсах программного обеспечения WEB-ориентированных систем. Обоснован выбор метрик для оценки качества информационного поиска. Сформулирована формальная постановка задачи классификации пользователей. Разработан метод коррекции профиля пользователя на основании содержимого информационного запроса, базируемый на модели нечеткой ожидаемой полезности. Построена упрощенная логическая модель данных задачи динамической классификации интересов пользователя в текущем сеансе. Разработан метод классификации пользователей.

Ключевые слова: *адаптивность, адаптируемость, интерфейс пользователя, классификация пользователей, нечеткое множество, Web-ориентированная система.*

Введение

Проектирование пользовательского интерфейса (ПИ) – это одновременно и искусство и наука, так как для оценки его качества используются как объективные, так и субъективные метрики. В настоящее время известные стандарты и показатели качества ПИ ориентированы на эмпирическую оценку уже созданного и применяемого ПИ, поэтому их применение на этапе проектирования ПИ практически невозможно. Между тем именно на этапе проектирования принимаются решения, изменение которых впоследствии методом проб и ошибок невероятно дорого и трудоемко. Специфика процесса разработки программного обеспечения (ПО) как инженерной дисциплины требует обеспечения качества архитектуры уже на этапе проектирования ПО, а не косвенно в уже внедренной системе.

WEB-ориентированные системы в настоящее время стали столь большими и сложными, что для их производства требуется участие слаженных команд разработчиков различных специальностей и квалификаций. Вложенные в их производство и освоение средства должны окупаться, поэтому такие системы должны существовать и применяться долгие годы, развиваясь от версии к версии, претерпевая на своем жизненном пути множество изменений, улучшая существующие и добавляя новые функции, корректируя и устраняя дефекты и ошибки. Длительный жизненный цикл предполагает способность WEB-ориентированных систем адаптироваться не только к изменению условий работы в новой среде, но и к изменяющимся требованиям пользователя.

В связи с требованиями постоянных изменений и с учетом сложности WEB-ориентированной системы возникает задача построения системы, позволяющей вносить изменения в сроки, ограниченные требованием эксплуатации. Такая система должна быть одновременно адаптируемой, то есть приспособленной к изменениям человеком в процессе жизненного цикла в соответствии с требованиями внешней среды, и адаптивной, то есть приспосабливающей свое поведение к каждому пользователю на основе нетривиальных выводов из информации об этом пользователе. Вопросы построения адаптивных систем, оценки их качества в существующих стандартах инженерии ПО изложены весьма в узкой постановке. Адаптивность рассматривается в контексте приспособленности к изменению апаратно-программного окружения, но не пользователя, что подчеркивает актуальность данного исследования.

В настоящее время модель пользователя остается недостаточно исследованной проблемой, прежде всего, потому, что она должна, с одной стороны, отражать субъективные особенности его поведения, с другой – специфику предметной области. В качестве предметной области в дальнейшем понимаются информационно-поисковые системы, адаптивность поведения которых предполагает навигационный выбор, поиск и предложение наиболее релевантных ссылок.

Целью работы является разработка метода классификации пользователей в адаптивных интерфейсах программного обеспечения WEB-ориентированных систем.

Для достижения цели необходимо решить следующие задачи:

– дать формальную постановку задачи классификации пользователей в адаптивных интерфейсах;

– обосновать выбор метрик для оценки качества информационного поиска;

– разработать метод классификации пользователей.

1. Формальная постановка задачи классификации пользователей в адаптивных интерфейсах программного обеспечения WEB-ориентированных систем

Метрики в информационных технологиях – это совокупность принципиально важных показателей, которые определяются и используются для оценки качества программных комплексов. Метрики определяют схему для выбора и специфицирования требований к качеству программных систем, а также для сопоставления возможностей различных программных продуктов.

Метрики внутреннего качества поиска. По аналогии со стандартом ISO 9126, определяющем шесть основных факторов внутреннего качества ПО, под внутренним качеством поиска будем понимать метрики, характеризующие близость в пространстве: текущего запроса пользователя; моделей пользователя и предметной области.

Решаемая задача относится к классу задач, в которых анализируемые объекты характеризуются многими разнородными признаками (количественными, качественными или смешанными).

При этом возможно существование экземпляров объектов, имеющих, в частности, и противоречивые описания, которые должны рассматриваться и анализироваться как единое целое, а свертка значений признаков или невозможна, или математически некорректна.

Принципиальными моментами кластерного анализа являются: выбор выражения, определяющего расстояние между объектами в признаковом пространстве; выбор алгоритма группирования объектов; разумная интерпретация сформированных групп. Выбор вида пространства и типа метрики зависит от свойств анализируемых объектов. Для рассмотренных выше многопризнаковых объектов наиболее адекватно метрическое пространство измеримых мультимножеств или множеств с повторяющимися элементами [1]. Кратность элементов – существенная особенность мультимножества, отличающая его от множества и позволяющая считать мультимножество качественно новым математическим понятием.

Под многопризнаковым объектом будем понимать объект, образованный алгебраическим произведением множеств:

$$Q = \langle S, D, T, V, W \rangle \times \langle \cup SP_{nj}, US_j \times UA_j \rangle, \qquad (1)$$

где $DM = \langle S, D, T, V, W \rangle$ – модель предметной области $\langle S, D, T, V, W \rangle$, состоящей из множеств рубрик – S, документов – D, понятий – T и отношений между ними. Основное назначение модели – хранить информацию об общем информационном поле предметной области;

$W = D \times T = \{ w_{i,j} \} / 0 < i \le \|D\|, 0 < j \le \|T\|$ - множество весов, характеризующих частоту использования понятий в документе;

$V = D \times S = \{ v_{k,l} \} / 0 < k \le \|D\|, 0 < l \le \|S\|$ - множество уверенностей, характеризующих субъективную оценку принадлежности документа d_k рубрике S_l.

Каждый документ с определенной степенью уверенности принадлежит одному из элементов множества рубрик S, то есть:

$$\forall k \cup \forall l, 0 < k \le \|D\|, 0 < l \le \|S\| : d_k \in s_l, \qquad (2)$$

поскольку в общем случае нет однозначного ответа относительно свойства документа.

Функция принадлежности указывает степень (или уровень) принадлежности документа определенной рубрике. То есть в рассмотрение вводится нечеткое множество $D \subset S$, отличающееся от обычного тем, что определяется как множество упорядоченных пар: $DS = \{ \mu DS(u) / u \}$, где $\mu D(u)$ – функция принадлежности, принимающая значение в интервале M = [0,1].

Помимо нечеткой принадлежности к рубрикам для любого документа характерна возможность четкого определения множества атрибутов A, таких как формат файла, язык, страна происхождения, которые используются далее в модели пользователя (МП). В целом документ – это его информационное содержание (контент) - C и атрибуты:

$$D = \langle C, A \rangle. \qquad (3)$$

Будем полагать, что множество рубрик S представляет собой направленный ациклический нечеткий граф $\tilde{G} = (S, \tilde{F})$, в котором определены множество вершин графа – S и нечеткое множество направленных ребер графа:

$$F = \{ \mu_F \langle s_i, s_j \rangle / \langle s_i, s_j \rangle \}, \ \langle s_i, s_j \rangle \in S^2, \qquad (4)$$

где $\mu_F \langle s_i, s_j \rangle$ – значение функции принадлежности для ребра $\langle s_i, s_j \rangle$ [2], при этом вершины являются инцидентными в том и только в том случае, если $\mu_F \langle s_i, s_j \rangle > 0$.

Ребро направленного нечеткого графа описывает принадлежность одной рубрики (темы) рубрике (теме) более высокого уровня.

МП – отображение модели предметной области в пространство интересов пользователя. Она представлена поисковым профилем пользователя, содержащим последовательность последних k поисковых запросов: $SP = \langle sp_1, sp_2, \ldots, sp_k \rangle$.

Множество поисковых профилей всех пользователей SP_i образует общее пространство запросов S, которое разбито на объединение непересекающихся множеств – кластеров UM_j:

$$S = \bigcup_{n=1}^{N} SP_n = \bigcup_{j=1}^{J} UM_j,$$

при условии:

$$\forall i,\ j \in 1\ldots J\ \ / UM_i \cap UM_j = \varnothing.$$

Каждый кластер UM_j описывается характеристиками:

$$UM_j = \left\langle \bigcup_{n=1}^{N} SP_{nj},\ US_j \times UA_j \right\rangle, \qquad (5)$$

где $\bigcup SP_{nj}$ – объединение поисковых профилей пользователей, отнесенных к j-му кластеру;

US_j – множество нечетких предпочтений пользователей, отнесенных к j-му кластеру, к определенным рубрикам документов;

UA_j – множество нечетких предпочтений пользователей, отнесенных к j-му кластеру, в пространстве атрибутов документов.

Нечеткое множество, определенное на носителе $UM_k \times US_j \times UA_j$, характеризует распределение интересов k-го пользователя по отдельным кластерам:

$$FUM = \left\{ x, \mu_{UM_{k\,j}}\left(UM_{k,\,j}\right) \middle| \ x \in UM_k \times US_j \times UA_j \right\},$$

где функция принадлежности $\mu_{UM_{k\,j}}\left(UM_{k,\,j}\right)$ указывает, насколько j-й кластер соответствует истории просмотров документов k-того пользователя. При этом значение функции принадлежности $\mu_{UM_{k\,j}}\left(UM_{k,\,j}\right)$, близкое к 0, свидетельствует об отсутствии интереса, а, близкое к 1, – об актуальности для k-го пользователя j-го кластера запросов.

Многопризнаковые объекты $A_i, i = 1,\ldots n$ принято представлять как векторы или кортежи $q_i = \left(q_{i1}^{el},\ldots,q_{im}^{em}\right)$ в пространстве $Q = Q_1 \times \ldots \times Q_m$, где $Q_s = \left\{q_s^{es}\right\}$ – непрерывная или дискретная шкала s-го признака, es=1÷hs, s=1,…,m, при этом одному и тому же объекту A_i может соответствовать не один, а несколько m-мерных векторов с разными значениями признаков. Подобные ситуации возникают, например, когда документ A_i востребован несколько раз одним или разными пользователями, или когда необходимо одновременно учесть m параметров объекта A_i, измеренных k различными способами.

В таких случаях объект A_i представляется в m-мерном пространстве Q не одним вектором q_i, а группой, состоящей из k векторов

$$q_i = \left\{q_i^{(1)},\ldots,q_i^{(k)}\right\}$$

вида $q_i^{(j)} = \left\{q_{i1}^{el(j)},\ldots,q_{im}^{em(j)}\right\}$, j=1,…,k, которая должна рассматриваться как единое целое. При этом, измеренные разными способами значения параметров, как и индивидуальные оценки, данные экспертами, могут быть похожими, различающимися и даже противоречивыми, что может приводить к несравнимости m-мерных векторов $q_i^{(j)}$, характеризующих один и тот же объект A_i.

Совокупность таких «составных» объектов имеет в пространстве Q сложную структуру трудную для анализа. Предлагаемая метрика для измерения расстояний между объектами в пространстве Q основана на формализме мультимножеств, позволяющем одновременно учесть различные комбинации значений количественных и качественных признаков и их многозначность.

Вместо прямого произведения m шкал признаков $Q = Q_1 \times \ldots \times Q_m$ введена обобщенная шкала признаков – множество $G = Q_1 \bigcup \ldots \bigcup Q_m$, состоящее из m групп признаков, объект A_i в таком символическом виде представляется как:

$$\begin{aligned} A_i = \{ & k_{A_i}\left(q_1^1\right) \circ q_1^1,\ldots, \\ & k_{A_i}\left(q_1^{h_1}\right) \circ q_1^{h_1},\ldots,k_{A_i}\left(q_m^1\right) \circ q_m^1,\ldots, \\ & k_{A_i}\left(q_m^{h_m}\right) \circ q_m^{hm}\}, \end{aligned}$$

где число $k_{A_i}\left(q_s^{es}\right)$ указывает, сколько раз признак $q_s^{es} \in Q_s$ встречается в описании объекта A_i, знак $\circ$ обозначает кратность вхождения признака q_s^{es}. Например, при многократном обращении нескольких пользователей к документу число $k_{A_i}\left(q_s^{es}\right)$ равно числу пользователей, обратившихся к документу A_i q_s^{es} - количество раз.

Над метрическими пространствами мультимножеств определены операции[3]:

– объединение

$$\begin{aligned} A \bigcup B = \{ & k_{A \cup B}(x) \circ x \mid k_{A \cup B}(x) \\ & = \max\left(k_A(x), k_B(x)\right)\}; \end{aligned}$$

– пересечение

$$\begin{aligned} A \bigcap B = \{ & k_{A \cap B}(x) \circ x \mid k_{A \cap B}(x) \\ & = \min\left(k_A(x), k_B(x)\right)\}; \end{aligned}$$

– сложение

$$A+B=\{k_{A+B}(x)\circ x \mid k_{A+B}(x) = k_A(x)+k_B(x)\};$$

– вычитание

$$A-B=\{k_{A-B}(x)\circ x \mid k_{A-B}(x) = k_A(x)-k_{A\cap B}(x)\};$$

– симметрическая разность

$$A\Delta B=\{k_{A\Delta B}(x)\circ x \mid k_{A\Delta B}(x) = |k_A(x)-k_B(x)|\};$$

– дополнение

$$\bar{A}=Z-A=\{k_{\bar{A}}(x)\circ x \mid k_{\bar{A}}(x) = k_Z(x)-k_A(x)\};$$

– умножение на число

$$t\bullet A=\{k_{t\bullet A}(x)\circ x \mid k_{t\bullet A}(x) = t\bullet k_A(x), t\in Z+\};$$

– умножение

$$A\bullet B=\{k_{A\bullet B}(x)\circ x \mid k_{A\bullet B}(x) = k_A(x)\bullet k_B(x)\};$$

– n-я степень

$$A^n=\{k_A n(x)\circ x \mid k_A n(x)=(k_A(x))^n\};$$

– прямое произведение

$$A\times B=\{k_{A\times B}\circ\langle x_i,x_j\rangle \mid k_{A\times B} = k_A(x_i)\cdot k_B(x_j), x_i\in A, x_j\in B\};$$

– прямая n-я степень

$$(\times A)^n=\{k_{(\times A)}n\circ\langle x_1,...,x_n\rangle \mid k_{(\times A)}n = k_A(x_1)\cdot...\cdot k_A(x_n), x_i\in A\},$$

где $\langle x_1,...,x_n\rangle$ – кортеж n элементов.

Мультимножество называется пустым $\varnothing$, если $k_\varnothing(x)=0$, и максимальным Z, если $k_z(x)=\max_{A\in A} k_A(x), \forall x\in G.$

Различные классы метрических пространств мультимножеств (A,d) задаются метриками (псевдометриками):

$$d_{1p}(A,B)=\left[m(A\Delta B)\right]^{1/p};$$

$$d_{2p}(A,B)=\left[m(A\Delta B)/m(Z)\right]^{1/p};$$

$$d_{3p}(A,B)=\left[m(A\Delta B)/m(A\cup B)\right]^{1/p},$$

где p – целое число, m – мера мультимножества, действительная неотрицательная функция, заданная на алгебре мультимножеств $L(Z)$.

Алгеброй мультимножеств называется семейство мультимножеств $L(Z)$, замкнутое относительно операций объединения, пересечения, сложения, вычитания и дополнения.

Максимальное мультимножество Z является единицей алгебры, а пустое мультимножество $\varnothing$ – нулем. Мера мультимножества m обладает свойствами:

– $m(\varnothing)=0$;

– сильной аддитивности

$$m\left(\sum_i A_i\right)=\sum_i m(A_i);$$

– слабой аддитивности

$$A_i\bigcap A_j=\varnothing;$$

– монотонности

$$m(A)\le m(B)\Leftrightarrow A\subseteq B;$$

– симметричности

$$m(A)+m(\bar{A})=m(Z);$$

– непрерывности

$$\lim_{i\to\infty} m(A_i)=m\left(\lim_{i\to\infty} A_i\right);$$

– эластичности

$$m(t\bullet A)=t\bullet m(A).$$

Меру мультимножества можно ввести различными способами, например, как мощность мультимножества $m(A)=|A|=\sum_i k_A(x_i)$ или как линейную комбинацию функций кратности $m(A)=\sum_i w_i k_A(x_i),\ w_i>0$. В этом случае метрики приобретают вид:

– полностью усредненная метрика

$$d_{2_p}(A,B)=\left(\sum_{x_i\in G} w'_i\,|k_A(x_i)-k_B(x_i)|\right)^{1/p}, \qquad (6)$$

где $w'_i=w_i/w'_i=w_i\Big/\sum_{j=1}^{h} w_j k_Z(x_j)$ характеризует различие между двумя мультимножествами A и B, отнесенное к расстоянию, максимально возможному в исходном пространстве;

– локально усредненная метрика

$$d_{3_p}(A,B)=\left(\frac{\sum_{x_i\in G} w_i\,|k_A(x_i)-k_B(x_i)|}{\sum_{x_i\in G} w_i\max[k_A(x_i),k_B(x_i)]}\right)^{1/p} \qquad (7)$$

задает различие, отнесенное к максимально возможной «общей части» $A\cup B$ только этих двух мультимножеств в исходном пространстве.

Функции $d_{2_p}(A,B)$ и $d_{3_p}(A,B)$ удовлетворяют условию нормировки $0\le d(A,B)\le 1$. Функция $d_{3_p}(A,B)$ не определена для $A=B=\varnothing$, поэтому по определению принимается $d_{3_p}(\varnothing,\varnothing)=0$.

Кластер – это теоретико-множественное объединение наиболее близких объектов. Рассмотренные операции над мультимножествами открывают новые возможности для группирования многопризнаковых объектов. Например, кластер X_t объектов может

быть получен как сумма $X_t = \sum_i A_i$, объединение $X_t = \bigcup_i A_i$ или пересечение $X_t = \bigcap_i A_i$ мультимножеств A_i, описывающих объекты A_i, либо как линейная комбинация различных мультимножеств вида $X_t = \sum_i t_i \bullet A_i$, $X_t = \bigcup_i t_i \bullet A_i$ или $X_t = \bigcap_i t_i \bullet A_i$, $t_i > 0$. Когда кластер X_t образуется в результате объединения или пересечения объектов A_i, происходит усиление лучших свойств (максимальных значений признаков x_j) или соответственно худших свойств (минимальных значений признаков x_j), присутствующих у отдельных членов группы. При сложении объектов агрегируются все свойства x_j всех членов кластера X_t.

Для выявления близких по свойствам объектов в алгоритмах кластерного анализа используются подходы:

– минимизировать различие (максимизировать сходство) между объектами внутри группы;

– максимизировать различие (минимизировать сходство) между группами объектов.

Метрики внешнего качества поиска. Существует два близких, но принципиально разных в контексте настоящего исследования понятия качества информационного поиска.

1. Релевантность - смысловое соответствие содержания найденных документов информационному запросу по конкретной теме.

2. Пертинентность - соответствие содержания документа реальной информационной потребности конкретного пользователя.

Метрики, перечисленные ниже, ориентированы на пертинентность.

Полнота R определяется как отношение количества правильно найденных при поиске документов к общему числу существующих по данному запросу документов:

$$R = a/(a+c), \qquad (8)$$

где a – количество правильно найденных документов;

c – количество правильных документов, которые система не смогла обнаружить.

Точность P определяется как отношение количества правильно найденных документов к общему количеству найденных документов.

$$P = a/(a+b), \qquad (9)$$

где b – количество неверно найденных документов.

Комбинированная F-мера позволяет учесть одновременно и полноту и точность:

$$F(P,R) = \frac{(\beta^2+1)P \cdot R}{\beta^2 \cdot P + R}, \qquad (10)$$

где β – параметр, задающий приоритет точности над полнотой.

Оценка пертинентности по определению субъективна, поэтому пользователю предпочтительно найти нужные ему документы в n –первых возвращаемых документах и нет необходимости просматривать всю коллекцию. Для учета этого фактора вводится функция точности на наборе первых n-возвращаемых поисковой машиной документов:

$$P(n) = a(n)/n, \qquad (11)$$

где $a(n)$ – зависимость количества правильно найденных документов от количества возвращаемых.

Формулировка задачи.

Дано:

– множество документов D (см. 1 – 4), которые после кластеризации объединены в тематически (содержательно) однородные группы - нечеткие кластеры KD;

– множество кластеров профилей всех пользователей UM_j (см. 5);

– модель актуального пользователя, его профиль SP_i;

– текущий запрос, как множество слов T для поиска, задающих информационную потребность пользователя $x = \langle T, W_T \rangle$,

– где T – множество ключевых слов, W_T – вес понятия (концепта) в модели конкретного пользователя;

– $D_j^{(n)}$ – j-я упорядоченная по релевантности последовательность документов длиной n из общего конечного множества документов D', полученного как результат решения задачи нечеткого поиска (ЗНП);

– $Da_j^{(n)}$ - конечное множество документов, возможно пустое, подмножество $D_j^{(n)}$, которые пользователь оценил пертинентными и запросил их просмотр.

Необходимо решить две задачи динамической классификации интересов пользователя в текущем сеансе (рис. 1), цель которых – коррекция оценок функций принадлежности профиля пользователя кластерам профилей всех пользователей $\mu_{UM_{k\,j}}(UM_{k,\,j})$. В первой задаче на основании информации содержащейся в формулировке запроса, как множестве слов T для поиска:

$$F_1 : FUM, T \to FUM. \qquad (12)$$

Во второй задаче входной информацией является $Da_j^{(n)}$ – множество пертинентных документов.

$$F_2 : FUM, Da_j^{(n)} \to FUM. \qquad (13)$$

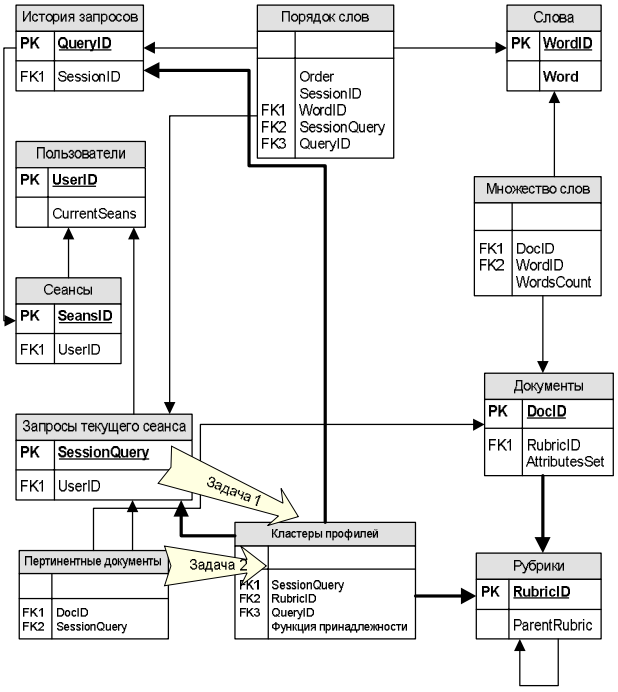

Рис. 1. Упрощенная логическая модель данных задачи динамической классификации интересов пользователя в текущем сеансе на (жирной линией выделены нечеткие отношения – соответствия)

2. Коррекция профиля пользователя на основании содержимого информационного запроса

Предполагается, что нечеткие кластеры профилей всех пользователей UM_j объединяют в себе пользователей с наиболее схожими потребностями. Для определения соответствующего кластера необходимо соотнести поисковый профиль конечного пользователя, состоящий из поисковых запросов текущего сеанса, со всеми кластерами.

Метод коррекции профиля пользователя на основании содержимого информационного запроса основан на модели нечеткой ожидаемой полезности. Данная модель используется в задачах индивидуального принятия решения, если наряду с классическим подходом нечеткой логики необходимо учесть теорию статистических решений [4] и теорию ожидаемой полезности [5]. Последняя предназначена для анализа решений, когда неопределенность обусловлена отсутствием объективной физической шкалы для оценки предпочтительности альтернатив. В этих случаях используется субъективная шкала полезности лица, принимающего решение (ЛПР). В реальных ситуациях исходы, соответствующие принятым решениям (состояниям системы), являются подчас неточными, что влечет за собой размытость соответствующих им оценок функции полезности. Размытый вариант ожидаемой полезности формулируется, например, в модели, с учетом случайных и нечетких составляющих неопределенности. Выбор происходит на основе максимизации нечеткой ожидаемой полезности:

$$\tilde{a} = \arg\max\left(\sum_{n=0}^{N} \tilde{p}_n F\left(s_n, a_j, b_k\right)\right), \qquad (14)$$

где $\tilde{p}_n$ – нечеткая вероятность состояния s_n из множества состояний S; $F\left(s_n, a_j, b_k\right)$ - функция полезности, определенная в пространстве множеств состояний S, альтернатив – A, критериев – B.

В качестве альтернативы n=0 будем понимать вариант предпочтительности сохранения исходного профиля пользователя

$$SP = \left\langle sp_1, sp_2, \ldots, sp_k \right\rangle.$$

Остальными альтернативами являются переход к кластерам профилей всех пользователей FUM.

Функция полезности ставит в соответствие множеству оценок – R-множество оценок полезности, то есть $F(R) = \left\{\left\langle R, \mu(R)\right\rangle \;\middle|\; \mu(R) \in [0,1]\right\}$. В качестве функции полезности использована локальная усредненная метрика для мультимножеств (см. 7), определяющая степень уверенности в принадлежности текущих интересов пользователя данному кластеру, при этом: мультимножество A – история запросов текущего сеанса пользователя; мультимножество B – объединение поисковых профилей пользователей, отнесенных к j-му кластеру $\cup SP_{nj}$.

Фактически, действительное значение функции полезности, определенной в пространстве множеств состояний S, альтернатив – A, критериев – B, задает линейный порядок на множестве исходов, выражающий прогноз предпочтительности кластера интересам пользователя в текущем сеансе. Возможны ситуации, когда будут отсутствовать похожие поисковые профили. В таком случае, поисковый профиль составляется на основе поисковых запросов в сеансах, с которыми просматриваемая пользователем страница была наиболее релевантной.

3. Коррекция профиля пользователя на основании известного множества пертинентных документов

Коррекция профиля пользователя на основании известного множества пертинентных документов сводится к вычислению композиции нечетких соответствий между:

– нечетким множеством пертинентных документов $Da_j^{(n)}$ и нечетким множеством рубрик – S;

– нечетким множеством рубрик – S и нечетким множеством кластеров профилей пользователей UMj.

Конечной целью вычисления композиции нечетких соответствий:

$$q = \left\langle Da_j^{(n)}, S, UM, DaS, SUM \right\rangle, \qquad (15)$$

где $DaS \subseteq Da_j^{(n)} \times S$, $SUM \subseteq S \times UM$ – нечеткие подмножества декартовых произведений, является ранжирование профилей пользователей UMj в соответствии с текущими интересами пользователя.

Вычисление функций принадлежности профилей пользователя производится по правилам нечеткого информационного графа.

Для обеспечения устойчивости процесса коррекции профиля пользователя на основании известного множества пертинентных документов применяется экспоненциальное сглаживание.

Метод экспоненциального сглаживания исходит из того, что истинная функция принадлежности профиля пользователя μ_{UM_j} меняется достаточно медленно по сравнению со случайной ошибкой, вызванной выбором нового документа во множестве пертинентных документов $Da_j^{(n)}$:

$$\mu_{UM_j}^{(n)} = \alpha \cdot y^{(n)} + (1-\alpha) \cdot \mu_{UM_j}^{(n-1)}, \qquad (16)$$

где $y^{(n)}$ – вычисленное значение функций принадлежности профиля пользователя.

При рекурсивном применении формулы каждое новое сглаженное значение (прогноз) вычисляется как взвешенное среднее текущего наблюдения и сглаженного ряда. Очевидно, результат сглаживания зависит от параметра α. Если α равен 1, то предыдущие наблюдения полностью игнорируются. Если α равен 0, то игнорируются текущие наблюдения. Значения α между 0 и 1 дают промежуточные результаты.

Эмпирические исследования показали, что простое экспоненциальное сглаживание часто дает достаточно точный прогноз.

Параметр сглаживания α подбирается адаптивно, то есть определяется такое значение α, для которого сумма квадратов (или средних квадратов) остатков (наблюдаемые значения минус прогнозы на шаг вперед) является минимальной.

Выводы

Для решения задачи классификации пользователей определены метрики, которые классифицированы на:

– метрики внутреннего качества, основанные на метрических свойствах мультимножества, образованного алгебраическим произведением моделей пользователя и предметной области

– метрики внешнего качества, характеризующего пертинентность результатов поиска.

Формальная постановка задачи динамической классификации интересов пользователя в текущем сеансе сведена к двум методам ее решения, основанных на обработке содержимого информационного запроса и множества пертинентных документов. В первом случае используются метрические свойства мультимножеств, во втором - вычисляется композиция нечетких соответствий между нечетким множествами пертинентных документов, рубрик и кластеров профилей пользователей UM_j. Для обеспечения устойчивости полученного решения использован метод экспоненциального сглаживания.

Дальнейшее развитие исследований предполагает проверку методов на известных наборах данных: выбор и обоснование метрик качества поиска, основанных на понятии пертинентности результатов поиска; разработку методов динамической классификации интересов пользователя в текущем сеансе.

Список литературы

1. Петровский А.Б. Пространства множеств и мультимножеств [Текст] /А.Б. Петровский. – М: Едиториал УРСС, 2003. – 248 с.

2. Мелихов А.Н. Ситуационные советующие системы с нечеткой логикой [Текст] / А.Н. Мелихов, Л.С. Берштейн, С.Я. Коровин. – М.: Наука, 1990. – 272 с.

3. Петровский А.Б. Метрические пространства мультимножеств [Текст] / А.Б. Петровский // Доклады Академии наук. – 1995. - Т. 344, № 2. - С. 175-177.

4. Ченцов А.Г. Элементы теории статистических решений (Байесовы и минимаксные решения) [Текст] / А.Г. Ченцов. - Екатеринбург: ГОУ ВПО УГТУ-УПИ, 2005. – 30 с.

5. Новоселов А.А. Математическое моделирование финансовых рисков: теория измерения [Текст] / А.А. Новоселов. - Новосибирск: Наука, 2001. - 99 с.

Поступила в редколлегию 21.03.2014

Рецензент: д-р техн. наук, проф. И.В. Шостак, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

КЛАСИФІКАЦІЯ КОРИСТУВАЧІВ У АДАПТИВНИХ ІНТЕРФЕЙСАХ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ WEB-ОРІЄНТОВАНИХ СИСТЕМ

Д.С. Негуриця, Т.Б. Шатовська

У статті виконано аналіз відомих методів динамічної класифікації користувачів в адаптивних інтерфейсах програмного забезпечення WEB-орієнтованих систем. Обґрунтовано вибір метрик для оцінки якості інформаційного пошуку . Сформульовано формальну постановку задачі класифікації користувачів . Розроблено метод корекції профілю користувача на підставі змісту інформаційного запиту, що базується на моделі нечіткої очікуваної корисності . Побудовано спрощену логічну модель даних задачі динамічної класифікації інтересів користувача в поточному сеансі. Розроблено метод класифікації користувачів.

Ключові слова: *адаптивність, адаптованість, інтерфейс користувача, класифікація користувачів, нечітка множина, Web-орієнтована система.*

USERS IN ADAPTIVE WEB-ORIENTED SYSTEM SOFTWARE INTERFACES CLASSIFICATION

D.S. Nehurytsia, T.B. Shatovska

In this article the familiar dynamic classification methods in adaptive user interface software WEB-oriented systems have been analyzed. Metrics choice for the information search quality estimation has been justified. A formal statement of the user's classification problem has been worded. A user profile correction method based on the information request content has been developed and based on the fuzzy expected utility model. A simplified logical data model of the user's interests dynamic classification in the current session task has been built. User's classification method has been developed.

Keywords: *agility, adaptability, user interface, users classification, user model, fuzzy set, Web-based system.*