Наука и Образование
МГТУ им. Н.Э. Баумана
Сетевое научное издание
ISSN 1994-0448

Наука и Образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2014. № 11. С. 571–584.

DOI: **10.7463/1114.0739214**

Представлена в редакцию: 15.11.2014

УДК 519.6

Оценка релевантности документов корпоративной онтологической базы знаний на основе их иерархической ролевой кластеризации

профессор, д.ф.-м.н. Карпенко А. П.[1,*], *apkarpenko@mail.ru
доцент, к.т.н. Трудоношин В. А.[1]

[1]МГТУ им. Н.Э. Баумана, Москва, Россия

Предполагаем, что метаданные документов базы знаний формируются на основе онтологии соответствующей предметной области, заданной в виде семантической сети. В отличие от предыдущих работ полагаем, что онтология предметной области, а тем самым и соответствующая семантическая сеть, иерархически кластеризована на основе ролей концептов. Предлагаем модель семантической сети базы знаний, использующей иерархическую ролевую кластеризацию концептов этой онтологии. Ставим задачу многокритериальной оценки релевантности документов базы знаний на основе их иерархической ролевой кластеризации. Предлагаем метод решения задачи, основанный на использовании оригинальных мер близости ролевых паттернов проектирования искомого документа и запроса.

Ключевые слова: корпоративная база знаний, онтология, семантическая сеть, релевантность

Введение

Корпоративная база знаний представляет собой совокупность большого числа разного рода слабоструктурированных документов, в которых с той или иной степенью детальности описаны прецеденты – некоторые ситуации и решения, которые были приняты в этих ситуациях. В системах поддержки принятия решений (СППР), которые используют такие базы знаний, поиск решения заключается в поиске в них наиболее подходящих прецедентов и соответствующих им документов [1].

Работа продолжает серию работ [1, 2], развивающих оригинальный подход к поиску решений в базах знаний прецедентов, в котором метаданные формируются на основе онтологии соответствующей предметной области, заданной в виде семантической сети. В отличие от указанных работ, в данной работе полагается, что онтология предметной области, а тем самым и соответствующая семантическая сеть, иерархически

кластеризована на основе ролей концептов. Онтологии, обладающие таким свойством, возникают, например, в медицине [3].

Как и в работах [1, 2], в данной публикации документы в базе знаний, а также поисковые запросы представляются в виде фреймов, которые называются паттернами проектирования и запроса соответственно. Слоты этих паттернов соответствуют ролям концептов используемой онтологии [4]. Указанные роли разбивают концепты онтологии, документа и запроса к базе знаний на кластеры. Предполагается, что по методике построения семантической сети документа, построены семантические сети указанных кластеров. Таким образом, поисковые образы документа и запроса представляются в виде совокупности семантических сетей, соответствующих слотам паттерна проектирования и паттерна запроса.

Современные поисковые системы основаны, как правило, на применении полнотекстового поиска, когда учитывается частота встречаемости терминов в документе, их средняя языковая частотность и так далее [5]. Известной альтернативой такому поиску является поиск по метаданным документа (авторы документа, его название, дата создания, тема и т.п.) [6]. Предлагаемый подход ставит целью повышение эффективности поиска решений в базах знаний прецедентов, основываясь не на регистрационных атрибутах документов, а на параметрах, характеризующих ситуацию принятия решения и само решение.

В первом разделе работы представляем модели семантических сетей онтологии и искомого документа. Во втором разделе описываем используемые паттерны проектирования документа и запроса. Третий раздел содержит постановку задачи оценки релевантности документа, как задачу целочисленной многокритериальной оптимизации. В четвертом разделе вводятся меры близости концептов и понятие их окрестностей. Пятый раздел представляет предлагаемые в работе меры релевантности документа базы знаний и запроса к этой базе. В заключении формулируем основные результаты работы и перспективы ее развития.

Полагаем далее, что $|A|$ - число элементов (мощность) счётного множества (набора) элементов A; a_q - q-й элемент этого множества (набора).

1. Модели семантических сетей онтологии и искомого документа

Пусть семантическую сеть рассматриваемой онтологии O определяет кортеж

$$SS = \langle C, R, P, E_c, E_r, E_p \rangle,$$

где приняты следующие обозначения:

$C = \{c_i\}$ - множество понятий семантической сети, для каждого из которых определена его *роль* $p(c_i) \in P$;

$R = \{r_j\}$ - множество отношений между понятиями набора C;

$P=\{p_k\}$ - множество ролей понятий C;

$E_c=\{e_{c,i}\}$, $E_r=\{e_{r,j}\}$, $E_p=\{e_{p,k}\}$ - значения *мер важности* понятий C, отношений R и ролей P соответственно.

Если понятия c_{i_1}, c_{i_2}, $i_1 \neq i_2$ в семантической сети SS связаны между собой некоторым отношением из числа отношений R, то говорим, что эти понятия *связаны информационно*.

Множество ролей P полагаем *линейно упорядоченным*, так что роль p_j предшествует роли p_{j+1}; $j \in [1:|P|-1]$.

Аналогично полагаем, что семантическая сеть искомого документа T рассматриваемой базы знаний определяет кортеж

$$SS^T = \left\langle C^T, R^T, P^T, E_c^T, E_r^T, E_p^T \right\rangle,$$

где $C^T \subseteq C$, $R^T \subseteq R$, $P^T \subseteq P$.

Сопоставляем семантической сети SS взвешенный ориентированный граф без контуров G, вершины которого соответствуют понятиям онтологии O, а дуги – информационным связям этих понятия между собой. Веса вершин графа G равны важностям w_i соответствующих понятий, а веса дуг $v_{i,j}$, $i, j \in [1:|O|]$ – важностям соответствующих отношений.

Полагаем, что семантическая сеть SS^T документа T включает в себя $|T|$ концептов и может быть представлена в виде аналогичного графу G взвешенного связного графа G^T, имеющего веса узлов w_i^T, веса ребер $v_{i,j}^T$, $i, j \in [1:|T|]$.

Исходим из того, что граф G может быть представлен в *ярусной форме*, удовлетворяющей следующим требованиям:

- число ярусов равно $|P|$;

- на верхнем ярусе находятся понятия с ролью p_1, на втором ярусе – понятия с ролями p_2, и так далее до яруса $|P|$, на котором находятся понятия с ролью $p_{|P|}$;

- любые два понятия, располагающиеся на одном ярусе, не связаны между собой информационно.

В указанных соглашениях и допущениях граф G^T также может быть представлен в аналогичной ярусной форме, число ярусов которой $|P^T|$ удовлетворяет условию $|P^T| \leq |P|$.

2. Паттерны проектирования искомого документа и запроса

Указанное в п. 1 ярусное представление графов G, G^T порождает ролевую кластеризацию семантических сетей SS, SS^T, так что множество концептов C оказывается разделенным на $|P|$ непересекающихся *ролевых кластеров* C_i, а множество концептов C^T документа T - на аналогичное число $|P^T|$ ролевых кластеров C_j^T; $i \in [1:|P|]$, $j \in [1:|P^T|]$. Кластерам C_i, C_i^T ставим в соответствие их семантические сети SS_i, SS_i^T и графы G_i, G_i^T. Обозначим $w_{i,p}$ - вес узла $c_{i,p}$ графа G_i; $v_{i,p,q}$ - вес ребра этого графа, связывающего его узлы $c_{i,p}, c_{i,q}$. Здесь $p,q \in [1:|C_i|]$, $p \neq q$; $|C_i|$ - число концептов в кластере C_i (равное числу узлов в графе G_i). Аналогичные обозначения $w_{i,p}^T$, $v_{i,p,q}^T$ введем для графа G_i^T [1].

Паттерн проектирования $A^T = \{A_i^T, i \in [1:|P^T|]\}$ документа T имеет $|P^T|$ слотов A_i^T и слот A_i^T соответствует роли p_i^T. Поисковый образ документа T представляет собой $|P^T|$ семантических сетей SS_i^T, формализованных в виде графов G_i^T; $i \in [1:|P^T|]$.

Введем еще следующие обозначения [1]:

$C^Q = \{c_i^Q, i \in [1:|C^Q|]\}$ - множество концептов запроса Q;

$\{C_i^Q,\ i \in [1:|P^Q|]\}$ - ролевые кластеры множества C^Q, где $|P^Q| \leq |P|$ - число ролей в запросе Q;

$\{SS_i^Q,\ i \in [1:|P^Q|]\}$ - совокупность семантических сетей запроса Q, где SS_i^Q - семантическая сеть ролевого кластера C_i^Q;

G_i^Q - граф семантической сети SS_i^Q; $i \in [1:|P^Q|]$;

$w_{i,p}^Q$ - вес узла $c_{i,p}^O$ графа G_i^Q; $i \in [1:|P^Q|]$;

$v_{i,p,q}^Q$ - вес ребра $(c_{i,p}^Q, c_{i,q}^Q)$ графа G_i^Q; $p,q \in [1:|P^Q|]$, $p \neq q$.

Таким образом, поисковый образ запроса Q представляет собой $|P^Q|$ семантических сетей SS_i^Q, формализованных в виде графов G_i^Q; $i \in [1:|P^Q|]$. Полагаем,

что образ запроса Q формирует паттерн $B^Q = \{B_i^Q, i \in [1:|P^Q|]\}$, который имеет $|P^Q|$ слотов B_i^Q [1].

3. Постановка задачи

Обозначим $\mathrm{M}_{i,j}(T,Q) = \{\mu_{i,j,k}(T,Q), k \in [1:|\mathrm{M}_{i,j}|]\}$ совокупность критериев релевантности, формализующих близость слотов A_i^T, B_j^Q паттернов проектирования документа T и запроса Q соответственно; $i \in [1:|P^T|]$, $j \in [1:|P^Q|]$. Полагаем, что большим значениям критерия $\mu_{i,j,k}(T,Q)$ соответствует большая релевантность документа T поисковому запросу Q.

Наборы всех критериев $\mathrm{M}_{i,j}(T,Q)$ образуют $(|P^T| \times |P^Q|)$ критериальную матрицу

$$\mathbf{M}(T,Q) = \{\mathrm{M}_{i,j}(T,Q), i \in [1:|P^T|],\ j \in [1:|P^Q|]\},$$

элементами которой являются $|\mathrm{M}_{1,j}|$-мерные векторные критерии $\mathrm{M}_{i,j}(T,Q)$. Таким образом, общее число критериев в матрице $\mathbf{M}(T,Q)$ равно

$$|\mathbf{M}| = \sum_{i=1}^{|P^T|}\sum_{j=1}^{|P^Q|}|\mathrm{M}_{i,j}|.$$

Ставим следующую дискретную задачу многокритериальной оптимизации (МКО). Среди всех документов $\{T\}$, имеющихся в базе знаний, найти документ T^*, который максимизирует матричный критерий релевантности $\mathbf{M}(T,Q)$:

$$\max_{T \in \{T\}} \mathbf{M}(T,Q) = \mathbf{M}(T^*,Q) = \mathbf{M}^*(Q). \qquad (2)$$

Критерии $\mu_{i,j,k}(T,Q)$, как правило, являются противоречивыми, так что документ T^*, максимизирующий некоторый критерий $\mu_{i_*,j_*,k_*}(T,Q)$ из числа этих критериев, в общем случае не доставляет максимум остальным указанным критериям. Поэтому запись (2) понимаем только в том смысле, что лицу, принимающему решения (ЛПР), желательна максимизации всех частных критериев оптимальности $\{\mu_{i,j,k}(T,Q)\}$ и что решением задачи является документ T^*.

Для предложенного в работе [2] адаптивного метода многокритериальной оценки релевантности число критериев $|\mathbf{M}|$ может быть неприемлемо большим. Поэтому с помощью какой-либо либо скалярной свертки перейдем от каждого из векторных

критериев $\mathrm{M}_{i,j}(T,Q)$ к скалярному критерию $\mu_{i,j}(T,Q)$. Например, для аддитивной скалярной свертки имеем

$$\mu_{i,j}(T,Q)=\sum_{k=1}^{\left|\mathrm{M}_{i,j}\right|}\alpha_{i,j,k}\ \mu_{i,j,k}(T,Q),\ i\in[1:\left|P^T\right|],\ j\in[1:\left|P^Q\right|],$$

где $\alpha_{i,j,k}\in[0;1]$ - вещественная константа, имеющая смысл относительного веса критерия $\mu_{i,j,k}(T,Q)$.

Таким образом, матрица векторных критериев $\mathrm{M}(T,Q)$ преобразуется в $(\left|P^T\right|\times\left|P^Q\right|)$ матрицу скалярных критериев. Сохраняем за этой матрицей прежнее обозначение. В результате дискретная МКО-задача (2) сводится к аналогичной задаче с матричным критерием оптимальности $\mathrm{M}(T,Q)$:

$$\max_{T\in\{T\}}\mathrm{M}(T,Q)=\mathrm{M}(T^*,Q)=\mathrm{M}^*(Q). \qquad (3)$$

4. Окрестность концепта семантической сети

Рассмотрим концепты $c_{i_1}, c_{i_n},\ i_1\neq i_n$ семантической сети SS, которой, напомним, соответствует граф G. Пусть $d(c_{i_1},c_{i_n})$ - *мера расстояния* между этими концептами. Величину $d(c_{i_1},c_{i_n})$ можно определить несколькими способами [7, 8]. Используем следующие меры.

Мера $d^1(c_{i_1},c_{i_n})$ - число дуг кратчайшего пути $c_{i_1},c_{i_2},...,c_{i_n}$ в графе G между вершинами, соответствующими указанным концептам, то есть

$$d^1(c_{i_1},c_{i_n})=i_n-i_1.$$

Заметим, что данная мера не учитывает важности концептов и отношений между ними.

Мера $d^2(c_{i_1},c_{i_n})$ учитывает важности $w_{i_1},w_{i_2},...,w_{i_n}$ концептов $c_{i_1},c_{i_2},...,c_{i_n}$ и отношений $v_{i_1,i_2},c_{i_2,i_3},...,c_{i_{n-1},i_n}$ между ними. Мера определяется выражением

$$d^2(c_{i_1},c_{i_n})=\alpha_{2,1}\sum_{k=i_1}^{i_n}w_k+\alpha_{2,2}\sum_{l=2}^{n}v_{i_{l-1},i_l}\ , \qquad (4)$$

где назначаемые ЛПР константы $\alpha_{2,1},\alpha_{2,2}\in[0;1]$ определяют веса важности концептов и отношений между ними.

Мера $d^3(c_{i_1},c_{i_n})$ учитывает еще и важность ролей концептов. Мера определяется выражением, аналогичным выражению (4):

$$d^3(c_{i_1}, c_{i_n}) = \alpha_{3,1} \sum_{k=i_1}^{i_n} w_k + \alpha_{3,2} \sum_{l=2}^{n} v_{i_{l-1}, i_l} + \alpha_{3,3} \sum_{m=1}^{i_n} e(c_m).$$

Здесь $e(c_m)$ - важность роли концепта c_m; $\alpha_{3,1}, \alpha_{3,2}, \alpha_{3,3} \in [0;1]$ - константы, определяющие веса важности концептов, отношений между ними и ролей концептов.

Введём в рассмотрение подграф $G_i(d)$ графа G, соответствующий концепту c_i. Подграф $G_i(d)$ включает в себя все концепты $C_i(d)$ графа G, расстояние от которых до концепта c_i в выбранной мере не превышает *d*. Набор ролей концептов $C_i(d)$ обозначаем $E_i(d)$; набор отношений, связывающих концепты этого набора - $R_i(d)$. Соответствующий фрагмент семантической сети $SS_i(d)$ называем *d-окрестностью* концепта c_i. Величина *d* называется *радиусом окрестности* $SS_i(d)$. Заметим, что окрестность $G_i(d^1), d^1 = 1$ представляет собой совокупность концептов, информационно связанных с концептом c_i.

4. Меры близости паттернов проектирования документа и запроса

Придадим частному критерию оптимальности $\mu_{i,j}(T, Q)$ смысл меры близости семантических сетей SS_i^T, SS_j^Q, соответствующих ролевым кластерам A_i^T, B_j^T паттернов проектирования документа и запроса соответственно.

Введем следующие обозначения:

$n_{i,j}$ - число концептов в ролевом кластере C_j^Q, которые содержатся в аналогичном кластере C_i^T (число верных концептов в запросе), то есть

$$n_{i,j} = \left| C_i^T \bigcap C_j^Q \right|;$$

$\overline{n}_{i,j}$ - число концептов в кластере C_j^Q, которые не входят в кластер C_i^T (число неверных концептов в запросе), то есть

$$\overline{n}_{i,j} = \left| C_j^Q \right| - n_{i,j}.$$

Мера $\mu_{i,j}^1(T, Q)$ не учитывает важность концептов и представляет собой аддитивную свертку взвешенных относительных чисел верных и неверных концептов в кластере C_j^Q:

$$\mu_{i,j}^1(T, Q) = \lambda_{1,1} \frac{n_{i,j}}{\left| C_i^T \right|} + \lambda_{1,2} \frac{\overline{n}_{i,j}}{\left| C_i^T \right|}. \qquad (5)$$

Здесь $\lambda_{1,1}, \lambda_{1,2} \in [0;1]$ - весовые коэффициенты, назначаемые ЛПР, исходя их своих предпочтений. В качестве методической основы определения значений этих коэффициентов может быть использован метод, предложенный в работе [9].

Заметим, что мера $\mu^1_{i,j}(T,Q)$ и другие аналогичные меры, представленные ниже, являются, по сути, многокритериальными, и указанные весовые коэффициенты определяет веса соответствующих частных критериев оптимальности [9]. Полагая все или некоторые весовые коэффициенты нулевыми, легко получить большое число производных мер.

Введем в рассмотрение величины

$$v^T_{i,j} = \sum_k w_k,\ c_k \in C^T_i \bigcap C^Q_j,$$

$$\bar{v}^T_{i,j} = \sum_k w_k,\ \ c_k \in C^Q_i,\ c_k \notin C^T_i \bigcap C^Q_j,$$

$$\mathrm{N}^T_i = \sum_k w_k,\ c_k \in C^T_i,$$

имеющие смысл суммарной важности верных концептов кластера C^Q_i, неверных концептов этого кластера, а также суммарной важности всех концептов указанного кластера соответственно.

Мера $\mu^2_{i,j}(T,Q)$ учитывает важность концептов и представляет собой аддитивную свертку взвешенных относительных чисел верных и неверных концептов в кластере C^Q_j с учетом важности этих концептов:

$$\mu^2_{i,j}(T,Q) = \lambda_{2,1}\frac{v_{i,j}}{\mathrm{N}^T_i} + \lambda_{2,2}\frac{\bar{v}_{i,j}}{\mathrm{N}^T_i},\ \lambda_{2,1}, \lambda_{2,2} \in [0;1]. \qquad (6)$$

Положим, что запрос Q включает в себя вектор $D = \{d_i, i \in [1:|P^Q|]\}$, компонента которого d_i имеет смысл радиуса окрестности каждого из концептов ролевого кластера C^Q_i, в которой следует искать концепты, релевантные концептам этого кластера.

Меры $\mu^3_{i,j}(T,Q,D)$, $\mu^4_{i,j}(T,Q,D)$ аналогичны мерам (5), (6) соответственно и основаны на предположении, что поиск концептов кластера C^T_i, релевантных концептам кластера C^Q_j, производится в d_i-окрестностях последних концептов.

Пусть $n_{i,j}(D)$ - число концептов кластера C^T_i, принадлежащих окрестностям всех концептов кластера C^Q_j (расширенное число верных концептов в запросе), то есть

$$n_{i,j}(D) = \left| \left(\bigcup_{k=1}^{|C_i|} C_k(d_i) \right) \bigcap C_j^Q \right|;$$

$\bar{n}_{i,j}(D)$ - число концептов в кластере C_j^Q, которые не входят в кластер C_i^T (расширенное число неверных концептов в запросе), то есть

$$\bar{n}_{i,j}(D) = \left| C_j^Q \right| - n_{i,j}(D).$$

В указанных обозначениях меру $\mu_{i,j}^3(T, Q, D)$ определяет выражение

$$\mu_{i,j}^3(T, Q, D) = \lambda_{3,1} \frac{n_{i,j}(D)}{\left| C_i^T \right|} + \lambda_{3,2} \frac{\bar{n}_{i,j}(D)}{\left| C_i^T \right|}, \ \lambda_{3,1}, \lambda_{3,2} \in [0; 1],$$

аналогичное выражению (5).

Введем еще следующие обозначения:

$$v_{i,j}^T(D) = \sum_k w_k, \quad c_k \in \left(\bigcup_{l=1}^{|C_i|} C_l(d_i) \right) \bigcap C_j^Q;$$

$$\bar{v}_{i,j}^T(D) = \sum_k w_k, \quad c_k \in \bigcup_{l=1}^{|C_i|} C_l(d_i), \ c_k \notin \left(\bigcup_{l=1}^{|C_i|} C_l(d_i) \right) \bigcap C_j^Q.$$

Здесь величины $v_{i,j}^T(D)$, $\bar{v}_{i,j}^T(D)$ имеют смысл суммарной важности концептов, принадлежащих и не принадлежащих d_i-окрестностям всех концептов кластера C_j^Q соответственно.

Мера $\mu_{i,j}^4(T, Q, D)$ аналогична мере (6) и определяется выражением

$$\mu_{i,j}^4(T, Q, D) = \lambda_{4,1} \frac{v_{i,j}(D)}{\mathrm{N}_i^T} + \lambda_{4,2} \frac{\bar{v}_{i,j}(D)}{\mathrm{N}_i^T}, \ \lambda_{4,1}, \lambda_{4,2} \in [0; 1].$$

Все представленные меры релевантности имеют, вообще говоря, разные знаки и масштаб. Поэтому в программных реализациях необходима нормировка этих мер по схеме

$$\mu_{i,j} = \frac{\mu_{i,j} - \mu_{i,j}^{\min}}{\mu_{i,j}^{\max} - \mu_{i,j}^{\min}} \in [0; 1],$$

где $\mu_{i,j}^{\min}$, $\mu_{i,j}^{\max}$ - минимально и максимально возможные значения меры $\mu_{i,j}$ [10].

Построение на основе нормированных метрик различных линейных и нелинейных бальных шкал оценок рассмотрено в работе [11].

Заключение

В работы предложена модель семантической сети онтологической базы знаний, использующей иерархическую ролевую кластеризацию концептов этой онтологии. Поставлена задача многокритериальной оценки релевантности документов корпоративной онтологической базы знаний на основе их иерархической ролевой кластеризации. Предложен метод решения задачи, основанный на использовании оригинальных мер близости ролевых паттернов проектирования искомого документа и запроса. Метод обладает высокой вычислительной сложностью и требует использования параллельных вычислительных систем [12].

Работа не снимает проблему лексической многозначности терминов [13]. Известен ряд методов решения данной задачи, например, методы, основанные на использовании Википедии [14].

В развитие работы планируется экспериментальная проверка эффективности предложенного метода.

Работа выполнена при поддержке гранта РФФИ 10-07-00222-а.

Список литературы

1. Карпенко А.П. Оценка релевантности документов онтологической базы знаний // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2010. № 9. Режим доступа: http://technomag.edu.ru/doc/157379.html (дата обращения 01.10.2014).
2. Карпенко А.П., Трудоношин В.А. Многокритериальная оценка релевантности документов корпоративной онтологической базы знаний на основе их ролевой кластеризации // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2013. № 11. С. 311-328. DOI: 10.7463/1113.0637857
3. Карпенко Д.С., Зарубина Т.В., Раузина С.Е., Богопольский Г.А., Тихонова Т.А., Глебова О.В. Система управления знаниями в медицинском вузе: взгляд на проблему, реалии, перспективы развития // Информационно-измерительные и управляющие системы. 2014. № 10. С. 10-18.
4. Норенков И.П. Интеллектуальные технологии на базе онтологий // Информационные технологии. 2010. № 1. С. 17-23.
5. Толчеев В.О. Методы выявления информационных признаков в задачах классификации текстовых документов // Информационные технологии. 2005. № 8. С. 14-21.
6. The Dublin Core® Metadata Initiative: website. Available at: http://dublincore.org/, accessed 01.10.2014.
7. Карпенко А.П., Галямова Е.В., Соколов Н.К. Методика контроля понятийных знаний субъекта обучения в обучающей системе // Наука и образование. МГТУ им. Н.Э.

Баумана. Электрон. журн. 2009. № 2. Режим доступа: http://technomag.edu.ru/doc/115086.html (дата обращения 01.10.2014).
8. Карпенко А.П., Соколов Н.К. Меры сложности семантической сети в обучающей системе // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2009. № 1. С. 50-66.
9. Галямова Е.В., Карпенко А.П., Соколов Н.К., Ягудаев Г.Г. Контроль понятийных знаний субъекта обучения в обучающей системе // Вестник МАДИ (ГТУ). 2009. № 2 (17). С. 82-86.
10. Карпенко А.П., Соколов Н.К. Оценка сложности семантической сети в обучающей системе // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2008. № 11. Режим доступа: http://technomag.edu.ru/doc/106658.html (дата обращения 01.10.2014).
11. Belous V.V., Bobrovsky A.V., Dobrjkov A.A., Karpenko A.P. , Smirnova E.V. Multi-criterion integral alternatives' estimation: mentally-structured approach to education // 2nd International Conference on Education and Education Management (EEM 2012), Hong Kong, China, September 4-5, 2012. Vol. 3. P. 215-224.
12. Лотов А.В., Поспелова И.И. Многокритериальные задачи принятия решений: учеб. пособие. М.: МАКС Пресс, 2008. 197 с.
13. Кобрицов Б.П. Методы снятия семантической многозначности // Научно-техническая информация. Сер. 2. 2004. № 2. С. 24-38.
14. Mihalcea R. Using Wikipedia for Automatic Word Sense Disambiguation // Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007), Rochester, April 2007. P. 196-203.

Science & Education of the Bauman MSTU
Electronic journal
ISSN 1994-0448

Science and Education of the Bauman MSTU, 2014, no. 11, pp. 571–584.

DOI: **10.7463/1114.0739214**

Received: 15.11.2014

Evaluating the Relevance of Corporate Ontological Knowledge Base Documents Using Their Hierarchical Role Clustering

A.P. Karpenko[1,*], **V.A. Trudonoshin**[1]

[*]apkarpenko@mail.ru

[1]Bauman Moscow State Technical University, Moscow, Russia

Keywords: corporate knowledge base, ontology, semantic network, relevance

The article considers a corporate knowledge base representing a set of the large number of various semi-structured documents, which describe to some detail precedents i.e. some situations and decisions made in these situations. The task is to find the decision in such knowledge base and search the most suitable precedents and their corresponding documents in it.

The work continues a series of the authors' works, which develop an original approach to the solution of this task. It is supposed that metadata of the knowledge base documents are based on the ontology of the corresponding subject domain specified as a semantic network. As opposed to the previous works it is necessary that ontology of subject domain, and thereby and corresponding semantic network, is hierarchically clustered based on the roles of concepts.

Documents in the knowledge base, and also search queries are presented as frames, which are called 'patterns of design and inquiry', respectively. Slots of these patterns correspond to roles of concepts of the used ontology. Above-noted roles break concepts of ontology, document, and request to the knowledge base into clusters. It is supposed that semantic networks of the abovementioned clusters are designed by a technique for creation of a semantic network of the document. Thus, search images of the document and inquiry are presented as a set of the semantic networks corresponding to slots of a pattern of design and a pattern of inquiry.

The work offers the model of a semantic network of the ontological knowledge base using a hierarchical role clustering of concepts of this ontology. The task for a multi-criteria evaluation of relevance of corporate ontological knowledge base documents using their hierarchical role clustering is set. The method is offered to solve the task using the original measures of proximity of role patterns of design of the required document and inquiry.

References

1. Karpenko A.P. Estimating document relevance in ontology knowledge base. *Nauka i obrazovanie MGTU im. N.E. Baumana = Science and Education of the Bauman MSTU*, 2010,

no. 9. Available at: http://technomag.edu.ru/doc/157379.html , accessed 01.10.2014. (in Russian).

2. Karpenko A.P., Trudonoshin V.A. Multi-criteria estimation of the relevancy of documents in the enterprise ontological knowledge base using thematic clusterization. *Nauka i obrazovanie MGTU im. N.E. Baumana = Science and Education of the Bauman MSTU*, 2013, no. 11, pp. 311-328. DOI: 10.7463/1113.0637857 (in Russian).
3. Karpenko D.S., Zarubina T.V., Rauzina S.E., Bogopol'skiy G.A., Tikhonova T.A., Glebova O.V. Knowledge management system in medical university: view on the problem, realities and perspectives of development. *Informatsionno-izmeritel'nye i upravlyayushchie sistemy = Information-measuring and Control Systems*, 2014, no. 10, pp. 10-18. (in Russian).
4. Norenkov I.P. Intellectual technologies on the base of ontologies. *Informatsionnye tekhnologii = Information Technologies*, 2010, no. 1, pp. 17-23. (in Russian).
5. Tolcheev V.O. Methods of feature selection in text categorization tasks. *Informatsionnye tekhnologii = Information Technologies*, 2005, no. 8, pp. 14-21. (in Russian).
6. The Dublin Core® Metadata Initiative: website. Available at: http://dublincore.org/, accessed 01.10.2014.
7. Karpenko A.P., Galyamova E.V., Sokolov N.K. The method of the student conceptual knowledge evaluation in the intellectual learning computer system. *Nauka i obrazovanie MGTU im. N.E. Baumana = Science and Education of the Bauman MSTU*, 2009, no. 2. Available at: http://technomag.edu.ru/doc/115086.html , accessed 01.10.2014.
8. Karpenko A.P., Sokolov N.K. Complexity Measures of Semantic Network of Learning System. *Vestnik MGTU. Ser. Priborostroenie = Herald of the Bauman MSTU. Ser. Instrument Engineering*, 2009, no. 1, pp. 50-66. (in Russian).
9. Galyamova E.V., Karpenko A.P., Sokolov N.K., Yagudaev G.G. Control of conceptual knowledge of the subject of training in training system. *Vestnik MADI (GTU)*, 2009, no. 2 (17), pp. 82-86. (in Russian).
10. Karpenko A.P., Sokolov N.K. Estimate of the complexity of semantic network into a tutoring system. *Nauka i obrazovanie MGTU im. N.E. Baumana = Science and Education of the Bauman MSTU*, 2008, no. 11. Available at: http://technomag.edu.ru/doc/106658.html , accessed 01.10.2014. (in Russian).
11. Belous V.V., Bobrovsky A.V., Dobrjkov A.A., Karpenko A.P. , Smirnova E.V. Multi-criterion integral alternatives' estimation: mentally-structured approach to education. *2nd International Conference on Education and Education Management (EEM 2012)*, Hong Kong, China, September 4-5, 2012, vol. 3, pp. 215-224.
12. Lotov A.V., Pospelova I.I. *Mnogokriterial'nye zadachi priniatiia reshenii* [Multicriterion problems of decision making]. Moscow, MAKS Press, 2008. 197 p. (in Russian).
13. Kobritsov B.P. Methods of removing semantic multivaluedness. *Nauchno-tekhnicheskaya informatsiya. Ser. 2*, 2004, no. 2, pp. 24-38. (in Russian).

14. Mihalcea R. Using Wikipedia for Automatic Word Sense Disambiguation. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, Rochester, April 2007, pp. 196-203.