НАУЧНОЕ ИЗДАНИЕ МГТУ ИМ. Н. Э. БАУМАНА
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС77 - 48211. Государственная регистрация №0421200025. ISSN 1994-0408
электронный научно-технический журнал

Многокритериальная оценка релевантности документов корпоративной онтологической базы знаний на основе их ролевой кластеризации

 # 11, ноябрь 2013
DOI: 10.7463/1113.0637857
Карпенко А. П., Трудоношин В. А.
УДК 519.6

Россия, МГТУ им. Н.Э. Баумана
apkarepnko@mail.ru
trudonoshin@mail.ru

Введение

Корпоративная база знаний представляет собой совокупность большого числа разного рода слабоструктурированных документов, в которых с той или иной степенью подробности 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].

Современные поисковые системы основаны, преимущественно, на применении полнотекстового поиска. При этом учитывается частота встречаемости терминов в документе, их средняя языковая частотность и так далее [2]. Альтернативой полнотекстовому поиску является поиск по метаданным, то есть поиск по атрибутам документов, содержащимся в их метаданных. Классический атрибутивный поиск основывается на использовании в качестве метаданных документа преимущественно его регистрационных атрибутов (авторы документа, название документа, дата создания, тема и т.п.) [3]. Эффективность поиска решений в базах знаний прецедентов можно повысить, основываясь не на регистрационных

атрибутах документов, а на параметрах, характеризующих ситуацию принятия решения и само решение.

Работа продолжает серию работ, использующих подход к поиску решений в базах знаний прецедентов, в котором метаданные формируются на основе онтологии соответствующей предметной области, заданной в виде семантической сети. Релевантность документов при этом может быть оценена с помощью значительного числа метрик, формализующих близость концептов, входящих в метаданные документа, и концептов поискового запроса [4 - 8].

О задаче поиска информации в общей постановке говорят в терминах модели поиска, которая включает в себя способ представления документов, способ представления поисковых запросов, вид критерия релевантности документов [9]. В данной работе документы в базе знаний, а также поисковые запросы представляются в виде фреймов, которые называются паттерном проектирования и паттерном запроса соответственно. Слоты этих паттернов соответствуют ролям концептов используемой онтологии (предметная область, объект, свойство, действие, задача и т.д.) [1].

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.

В пером разделе работы приводим постановку задачи многокритериальной оценки релевантности документов корпоративной онтологической базы знаний. Во втором разделе даем определения множества и фронта Парето поставленной задачи. Третий раздел содержит обзор известных классов методов многокритериальной оптимизации и обоснование выбора класса адаптивных методов многокритериальной оптимизации. В четвертом разделе представляем предлагаемый адаптивный

метод многокритериальной оценки релевантности. В заключении формулируем основные результаты работы и перспективы ее развития.

1. Постановка задачи

Положим, что, например, по методике, предложенной в работе [10], семантическая сеть $S(O)$ рассматриваемой онтологии O построена в виде взвешенного связного графа $G(O)$ с весами узлов w_i^O и весами ребер $v_{i,j}^O$; $i, j \in [1:|O|]$. Пусть аналогично определена семантическая сеть $S(T)$ рассматриваемого документа T в виде взвешенного связного графа $G(T)$, имеющего веса узлов w_i^T , веса ребер $v_{i,j}^T$ и «расстояние» между узлами $l_{i,j}^T$; $i, j \in [1:|T|]$, $|T| < |O|$. Здесь и далее запись вида $|\Omega|$, где Ω - некоторое множество или вектор, означает мощность этого множества или размерность вектора соответственно.

Тем или иным способом, произведена ролевая кластеризация семантических сетей $S(O)$, $S(T)$, так что множество концептов $C(O)$ разделено на $|D|$ непересекающихся «ролевых» кластеров D_i^O, а множество концептов C^T документа T - на такое же число ролевых кластеров D_i^T; $i \in [1:|D|$. Кластерам D_i^O, D_i^T ставим в соответствие их семантические сети S_i^O, S_i^T и графы G_i^O, G_i^T. Обозначим $w_{i,p}^O$ - вес узла $c_{i,p}$ графа G_i^O, $v_{i,p,q}^O$ - вес ребра этого графа, связывающего его узлы $c_{i,p}, c_{i,q}$. Здесь $p, q \in [1:|D_i^O|]$, $p \neq q$; $|D_i^O|$ - число концептов в кластере D_i^O (равное числу узлов в графе G_i^O). Аналогичные обозначения $w_{i,p}^T$, $v_{i,p,q}^T$ введем для графа G_i^T [10].

Пусть паттерн проектирования $A(T) = \{A_i(T), i \in [1:|D|]\}$ документа T имеет $|D|$ слотов $A_i(T)$ и слот $A_i(T)$ соответствует роли ω_i. Поисковый образ документа T представляет собой $|D|$ семантических сетей S_i^T ,

формализованных в виде графов G_i^T; $i \in [1:|D|]$. Не ограничивая общности рассуждений, полагаем, что поисковый образ запроса Q формирует паттерн $B(Q) = \{B_i(Q), i \in [1:|D|]\}$, который также имеет $|D|$ слотов $B_i(Q) = B_i$ [10].

Введем следующие обозначения: C^Q - множество концептов запроса Q; $|Q|$ - число концептов во множестве C^Q; D_i^Q - ролевые кластеры множества C^Q, $i \in [1:|D|]$; C_i^Q - множество концептов кластера D_i^Q; $|D_i^Q|$ - число концептов в кластере D_i^Q; S_i^Q - семантическая сеть кластера D_i^Q; G_i^Q - граф семантической сети S_i^Q; $w_{i,p}^Q$ - вес узла $c_{i,p}^O$ графа G_i^Q; $v_{i,p,q}^Q$ - вес ребра $(c_{i,p}^Q, c_{i,q}^Q)$ графа G_i^Q. Здесь $p, q \in [1:|D_i^Q|]$, $p \neq q$. Таким образом, поисковый образ запроса Q представляет собой $|D|$ семантических сетей S_i^Q, формализованных в виде графов G_i^Q; $i \in [1:|D|]$ [10].

Обозначим $R(T, Q) = \{r_j(T, Q), j \in [1:|R|]$ совокупность критериев релевантности, формализующих близость семантических сетей S_i^T поискового образа документа T и семантических сетей S_i^Q запроса Q; $i \in [1:|D|]$. Полагаем, что большим значениям критерия $r_j(T, Q)$ соответствует большая релевантность документа T поисковому запросу Q.

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

$$\max_{T \in \{T\}} R(T, Q) = R(T^*, Q) = R^*(Q). \quad (1)$$

Поскольку речь идет о фиксированном запросе Q, в дальнейших записях символ Q будем опускать.

2. Множество и фронт Парето задачи многокритериальной оценки релевантности

Критерии $r_j(T)$, $j \in [1:|R|]$, как правило, являются противоречивыми, так что документ T^*, максимизирующий критерий $r_j(T)$, в общем случае не доставляет максимум остальным указанным критериям. Поэтому запись (1) следует понимать только в том смысле, что лицу, принимающему решения (ЛПР), желательна максимизация всех критериев $r_j(T)$, $j \in [1:|R|]$.

Критериальная вектор-функция $R(T)$ выполняет отображение множества $\{T\}$ в некоторое множество $\{R\}$ критериального пространства, которое называется *множеством достижимости* задачи (1). Введем на множествах $\{R\}$, $\{T\}$ отношение доминирования.

Вектор $R_1 = R(T_1) \in \{R\}$ *доминирует* вектор $R_2 = R(T_2) \in \{R\}$, что записываем в виде $R_1 \succeq R_2$, если среди равенств и неравенств $r_k(T_1) \geq r_k(T_2)$, $k \in [1:|R|]$ имеется, хотя бы одно строгое.

Документ $T_1 \in \{T\}$ *доминирует* документ $T_2 \in \{T\}$, то есть $T_1 \trianglerighteq T_2$, если $R(T_1) \succeq R(T_2)$.

Выделим из множества $\{R\}$ подмножество точек P_R^* - *фронт Парето* МКО-задачи (1), которые не доминируются другими точками множества $\{R\}$ и среди которых нет доминирующих друг друга. Множество $P_T^* \in \{T\}$, соответствующее множеству P_R^*, называют *множеством Парето* указанной МКО-задачи. Таким образом, если $T \in P_T^*$, то $R(T) \in P_R^*$.

Множество Парето и фронт Парето занимают в теории многокритериальной оптимизации исключительное место. Это обусловлено тем обстоятельством, что согласно известному *принципу Эджворта-Парето*, при «разумном» поведении ЛПР выбор решения следует производить на множестве Парето.

Роль множества Парето при решении задач МКО определяет также следующая теорема [11].

Теорема. Если для некоторых весовых множителей $\lambda_j > 0$, $j \in [1:|R|]$ имеет место равенство

$$\max_{T \in \{T\}} \sum_{j=1}^{|R|} \lambda_j \, r_j(T) = \sum_{j=1}^{|R|} \lambda_j \, r_j(T^*), \tag{2}$$

то вектор T^* оптимален по Парето, то есть $T^* \in P_T^*$.

Теорема показывает, что выбор определенной точки из множества Парето эквивалентен указанию весов каждого из частных критериев оптимальности. На этом факте основано большое число приближенных алгоритмов решения задач МКО.

Заметим, что теорема задает лишь *необходимое условие* оптимальности по Парето вектора $T^* \in \{T\}$. Другими словами, из того факта, что точка T^* принадлежит множеству Парето, не следует, что эта точка обязательно удовлетворяет условию (2).

Постановка МКО-задачи (1) фиксирует лишь множество допустимых значений $\{T\}$ и вектор критериальных функций $R(T) = (r_1(T), r_2(T), ..., r_{|R|}(T))$. Как правило, этой информации недостаточно для однозначного решения указанной задачи. Данная информация позволяет лишь выделить соответствующее множество Парето. Поэтому часто говорят, что решением МКО-задачи в постановке (1) является множество Парето этой задачи. Для однозначного решения задачи нужна дополнительная информация, которая в той или иной форме может быть предоставлена только ЛПР.

3. Обзор методов многокритериальной оптимизации

Методы решения задачи МКО чрезвычайно разнообразны. Существует несколько способов классификации этих методов. Используем в качестве основы классификацию, предложенную в работе [11], и выделим следующие классы методов решения МКО-задачи:

- априорные методы;

- апостериорные методы;

- адаптивные методы;

- методы Парето-аппроксимации.

Методы каждого из указанных классов имеет свои достоинства и ни один из них не свободен от недостатков, наличие которых не позволяет признать методы этого класса универсальными. Эти же классы методов в значительной мере переплетаются друг с другом так, что не всегда МКО-метод удается однозначно отнести к тому или иному классу. Так, в настоящее время развивается класс адаптивных эволюционных методов, которые основаны на многократном построении некоторых фрагментов множества Парето. Примером алгоритмов этого класса служит алгоритм *MOEA/D* (*Multiobjective Evolutionary Algorithm based on Decomposition*) [12].

Общей идеей методов решения МКО-задачи является сужение множества $\{T\}$ вплоть до одной или немногих точек. Построение множества Парето или его некоторой аппроксимации можно интерпретировать как часть этого пути.

Априорные методы предполагают формализацию дополнительной информации о предпочтениях ЛПР до начала решения задачи, то есть априори. Чаще всего эту информацию представляют в форме, позволяющей свести многокритериальную задачу к однокритериальной задаче оптимизации некоторой скалярной функции. Наиболее известным алгоритмом этого класса является *алгоритм скалярной свертки*. В данном случае указанную скалярную функцию образует, например, взвешенная сумма частных критериев (аддитивная скалярная свертка) вида (2), где λ_k - априори назначаемые ЛПР веса частных критериев оптимальности. Недостатком аддитивной свертки является невозможность с ее помощью получить решения, принадлежащие невыпуклым частям фронта Парето МКО-задачи. Если вместо свертки вида (2) использовать известную *свертку*

Джоффриона на основе *лексикографического упорядочения* [1], то этот метод можно использовать и в случае невыпуклого фронта Парето.

В силу простоты, априорные методы чаще всего использует в вычислительной практике решения МКО-задач. Недостатком этих методов является то обстоятельство, что в общем случае относительная важность частных критериев оптимальности может быть определена только в процессе многократного решения задачи (1), так что, в конечном счете, методы этого класса оказываются апостериорными.

Апостериорные методы предполагают внесение ЛПР в МКО-систему (программную систему, реализующую соответствующий апостериорный метод) дополнительной информации о своих предпочтениях в ходе решения задачи, то есть апостериори. При этом обычно дополнительную информацию, как и в априорных алгоритмах, формализуют в виде некоторой скалярной целевой функции. Примером апостериорного метода может служить известный *метод последовательных уступок* [11].

Адаптивные методы включают в себя некоторую совокупность итераций, каждая из которых содержит фазу анализа, выполняемого ЛПР, и фазу расчетов, выполняемых МКО-системой. По характеру информации, получаемой этой системой от ЛПР на фазе анализа, выделяют три класса адаптивных методов:

- методы, в которых ЛПР непосредственно назначает весовые коэффициенты частных критериальных функций;
- методы, в которых ЛПР накладывает ограничения на значения этих функций;
- методы, которые предполагают только оценку ЛПР альтернатив, предлагаемых ему МКО-системой.

В последнем случае может производиться бальная оценка альтернатив (например, в терминах «отлично», «хорошо», «удовлетворительно» и т.д.) либо парное сравнение альтернатив между собой (например, в терминах «лучше», «хуже», «одинаково»).

В конечном счете, априорные, апостериорные и адаптивные методы сводят решение МКО-задачи к однокритериальной задаче глобальной (в нашем случае, дискретной) оптимизации.

Алгоритмы Парето-аппроксимации не предполагают формализации в той или иной форме дополнительной информации о предпочтениях ЛПР. Алгоритмы этого класса предполагают построение тем или иным способом аппроксимации множества и фронта Парето МКО-задачи и предъявление полученных результатов ЛПР для их неформального анализа и выбора одного из решений. В случае если полученные множества содержат большое число точек, их анализ ЛПР может быть затруднительным.

4. Адаптивный метод многокритериальной оценки релевантности

Авторы предлагают модификацию адаптивного метода *PREF* [13] для решения задачи многокритериальной оценки релевантности документов (1). Метод предполагает бальную оценка ЛПР альтернатив, предлагаемых ему МКО-системой.

Положим, что частные критерии оптимальности $r_1(T), r_2(T),..., r_{|R|}(T)$ тем или иным образом нормализованы [4], так что $r_i(T) \in [0;1]$ для любого $T \in \{T\}$. Рассматриваем решение задачи (1) методом скалярной свертки. Способ свертки не фиксируем – это может быть аддитивная свертка, мультипликативная свертка, свертка Джоффриона и другие свертки [11]. Обозначим операцию свертки $\varphi(T, \Lambda)$, где $\Lambda \in D_\Lambda$ - вектор весовых множителей, $D_\Lambda = \{\lambda_i \mid \lambda_i \geq 0, \sum_i \lambda_i = 1, \quad i \in [1:|R|]\}$ - множество допустимых значений этого вектора.

При каждом фиксированном векторе $\Lambda \in D_\Lambda$ метод скалярной свертки сводит решение задачи (1) к решению однокритериальной задачи глобальной условной дискретной оптимизации

$$\max_T \varphi(T, \Lambda) = \varphi(T^*, \Lambda), \; T \in \{T\}. \qquad (3)$$

В силу счетности множества $\{T\}$ решение этой задачи существует.

Если при каждом $\Lambda \in D_\Lambda$ решение задачи (3) единственно (а при численной реализации это всегда можно обеспечить), то это решение ставит в соответствие каждому из допустимых векторов Λ единственный вектор T^* и соответствующие значения частных критериев оптимальности $r_1(T^*), r_2(T^*),..., r_{|R|}(T^*)$. Данное обстоятельство позволяет полагать, что в этом случае функция предпочтений ЛПР ψ определена не на множестве $\{T\}$, а на множестве D_Λ:

$$\psi : \Lambda \to \mathbf{R}.$$

В результате МКО-задача (1) сводится к задаче выбора вектора Λ^* такого, что

$$\max_{\Lambda} \psi(\Lambda) = \psi(\Lambda^*),\ \Lambda \in D_\Lambda. \qquad (4)$$

Если используется аддитивная свертка и множество достижимости $\{R\}$ является выпуклым, то выражение (3) задает взаимно однозначное отображение множества D_Λ на множество P_R^*. При этом для любого $\Lambda \in D_\Lambda$ вектор T^*, являющийся решением задачи (3), принадлежит множеству Парето P_T^*. Если вместо аддитивной свертки используется свертка Джоффриона, то для получения того же результата не требуется выпуклость множества достижимости [11].

Величину ψ считаем лингвистической переменной со значениями представленными в таблице 1, где $\overset{\circ}{\psi}$ - ядро нечеткой переменной ψ [14].

Таблица 1. Допустимые значения функции предпочтений ЛПР, как лингвистической переменной

Функция предпочтений ψ	$\overset{\circ}{\psi}$
”Очень-очень плохо”	1
”Очень плохо”	2
”Плохо”	3

”Не совсем удовлетворительно”	4
”Удовлетворительно”	5
”Не совсем хорошо”	6
”Хорошо”	7
”Очень хорошо”	8
”Отлично”	9

В результате МКО-задача (1) сводится к задаче отыскания вектора Λ^*, обеспечивающего максимальное значение дискретной функции $\overset{\circ}{\psi}(\Lambda)$:

$$\max_{\Lambda} \overset{\circ}{\psi}\Lambda) = \overset{\circ}{\psi}(\Lambda^*) = \overset{\circ}{\psi}^*, \ \Lambda \in D_\Lambda. \qquad (5)$$

Общая схема предлагаемого метода решения задачи (1) является итерационной и состоит из следующих основных этапов.

Этап «разгона» метода. МКО-система некоторым образом (например, случайно) последовательно генерирует k векторов $\Lambda_1, \Lambda_2, ..., \Lambda_k$ и для каждого из этих векторов выполняет следующие действия:

1) решает задачу дискретной глобальной оптимизации

$$\max_{T} \varphi(T, \Lambda_i) = \varphi(T_i^*, \Lambda_i), \ T \in \{T\}, \ i \in [1:k]; \qquad (6)$$

2) предъявляет ЛПР найденный документ T_i^*, а также соответствующие значение всех частных критериев оптимальности $r_1(T_i^*), r_2(T_i^*), ..., r_{|R|}(T_i^*)$;

3) ЛПР оценивает эти данные и вводит в МКО-систему соответствующее значение своей функции предпочтений $\overset{\circ}{\psi}(\Lambda_i)$.

Первый этап. На основе всех имеющихся в МКО-системе значений $\Lambda_1, \Lambda_2, ..., \Lambda_k$ вектора Λ и соответствующих оценок функции предпочтений $\psi(\Lambda_1), \psi(\Lambda_2), ..., \psi(\Lambda_k)$ МКО-система выполняет следующие действия.

1) Строит функцию $\tilde{\psi}_1(\Lambda)$, аппроксимирующую функцию $\psi(\Lambda)$ в окрестности точек $\Lambda_1, \Lambda_2, ..., \Lambda_k$.

2) Отыскивает максимум функции $\tilde{\psi}_1(\Lambda)$ – решает задачу

$$\max_{\Lambda} \tilde{\psi}_1(\Lambda) = \tilde{\psi}(\Lambda_1^*), \ \Lambda \in D_\Lambda.$$

3) С найденным вектором Λ_1^* решает задачу вида (6) – находит соответствующий документ и значения частных критериев оптимальности, а затем предъявляет их ЛПР. ЛПР оценивает указанные данные и вводит в систему соответствующее значение своей функции предпочтений $\psi(\Lambda_1^*)$.

Второй этап. На основе всех имеющихся в системе значений $\Lambda_1, \Lambda_2, ..., \Lambda_k, \Lambda_1^*$ вектора Λ и соответствующих оценок функции предпочтений $\psi(\Lambda_1), \psi(\Lambda_2), ..., \psi(\Lambda_k), \psi(\Lambda_1^*)$ МКО-система выполняет аппроксимацию функции $\psi(\Lambda)$ в окрестности точек $\Lambda_1, \Lambda_2, ..., \Lambda_k, \Lambda_1^*$ - строит функцию $\tilde{\psi}_2(\Lambda)$ по схеме первого этапа и так далее до тех пор, пока ЛПР не примет решение о прекращении вычислений.

В работе [15] для аппроксимации функции предпочтений ЛПР предложено использовать нейронные сети, аппарат нечеткой логики, а также нейро-нечеткие системы. В современной вычислительной практике эти средства широко применяют для решения плохо формализуемых задач, к которым относится задача аппроксимации функции предпочтений.

Результаты исследования эффективности указанных методов аппроксимации функции предпочтений ЛПР ожидаемо показывают их близкую эффективность. В силу относительной простоты реализации и сравнительно невысокой вычислительной сложности останавливаемся на аппроксимации функции предпочтений ЛПР с помощью нейронных сетей. По сравнению с классической полиномиальной аппроксимации на основе регрессионных планов, нейросетевая аппроксимация имеет следующие преимущества:

- нейронные сети способны моделировать широкий класс функциональных зависимостей, при использовании же полиномов класс функций, как правило, должен быть задан;

- для нейронных сетей существует эффективный способ настройки их параметров.

Заключение

Задача оценки релевантности документа представляет собой, по сути, задачу многокритериальной оптимизации. До настоящего времени эта задача рассматривалась как однокритериальная или как многокритериальная, но сводящаяся к многокритериальной методом аддитивной скалярной свертки. Этот метод прост в реализации, но далеко не всегда является эффективным. В частности, в общем случае этот метод не гарантирует отыскание всех паретовских точек (если фронт Парето задачи не является выпуклым). На необходимость использования иных методов решения задачи многокритериальной оценки релевантности указывалось еще в работе [10].

В работе предложен адаптивный метод многокритериальной оценки релевантности документов, основанный на ролевой кластеризации документов в корпоративной онтологической базе знаний и аппроксимации функции предпочтений ЛПР.

Даже однокритериальный вариант этого метода обладает высокой вычислительной сложностью и требует использования параллельных вычислительных систем.[10]. Тем более использование этих систем необходимо при реализации многокритериального варианта метода.

Одной из проблем, которая возникает при использовании рассмотренного подхода к определению релевантности документов (как в однокритериальном, так и в многокритериальном вариантах), является проблема лексической многозначности терминов. Правильное значение многозначного слова может быть установлено только путем анализа контекста, в котором это слово упоминается. Известен ряд методов решения данной задачи, например, методы, основанные на использовании Википедии [16].

В развитие работы планируется экспериментальная проверка эффективности предложенного подхода.

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

Список литературы

1. Норенков И.П. Интеллектуальные технологии на базе онтологий // Информационные технологии. 2010. № 1. С. 17-23.
2. Толчеев В.О. Методы выявления информационных признаков в задачах классификации текстовых документов // Информационные технологии. 2005. № 8. С. 14-21.
3. The Dublin Core® Metadata Initiative. Режим доступа: http://dublincore.org/ (дата обращения 01.10.2013).
4. 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 01.10.2013).
5. Карпенко А.П., Соколов Н.К. Расширенная семантическая сеть обучающей системы и оценка ее сложности // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2008. № 12. Режим доступа: http://technomag.edu.ru/doc/111716.html (дата обращения 01.10.2013).
6. Галямова Е.В., Карпенко А.П., 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 01.10.2013).
7. Карпенко А.П., Соколов Н.К. Меры сложности семантической сети в обучающей системе // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2009. № 1 (74). С. 50-66.
8. 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.
9. Когаловский М.Р. Перспективные технологии информационных систем. М.: ДМК Пресс; Компания АйТи, 2003. 288 с.

10. 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 01.10.2013).

11. Лотов А.В., Поспелова И.И. Многокритериальные задачи принятия решений: учеб. пособие. М.: МАКС Пресс, 2008. 197 с.

12. Zhang Q., Li H. MOEA/D: A multiobjective evolutionary algorithm based on decomposition // IEEE Transactions on Evolutionary Computation. 2007. Vol. 11, no. 6. P. 712-731. DOI: 10.1109/TEVC.2007.892759

13. Карпенко А.П., Мухлисуллина Д.Т., Овчинников В.А. Нейросетевая аппроксимация функции предпочтений лица, принимающего решения, в задаче многокритериальной оптимизации // Информационные технологии. 2010. № 10. С. 2-9.

14. Мухлисуллина Д.Т., Моор Д.А., Карпенко А.П. Многокритериальная оптимизация на основе нечеткой аппроксимации функции предпочтений лица, принимающего решения // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2010. № 1. Режим доступа: http://technomag.edu.ru/doc/135375.html (дата обращения 01.10.2013).

15. Карпенко А.П., Федорук В.Г. Аппроксимация функции предпочтений лица, принимающего решения, в задаче многокритериальной оптимизации. 3. Методы на основе нейронных сетей и нечеткой логики // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2008. № 4. Режим доступа: http://technomag.edu.ru/doc/86335.html (дата обращения 01.10.2013).

16. Mihalcea R. Using Wikipedia for Automatic Word Sense Disambiguation // Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007). Rochester, NY, USA, April 2007. P. 196-203.

SCIENTIFIC PERIODICAL OF THE BAUMAN MSTU
SCIENCE and EDUCATION
EL № FS77 - 48211. №0421200025. ISSN 1994-0408
electronic scientific and technical journal

Multi-criteria estimation of the relevancy of documents in the enterprise ontological knowledge base using thematic clusterization

 # 11, November 2013
DOI: 10.7463/1113.0637857

Karpenko A.P., Trusonoshin V. A.
Bauman Moscow State Technical University, 105005, Moscow, Russian Federation
apkarepnko@mail.ru
trudonoshin@mail.ru

This work is part of the studies in the process of development of design methods for ontological knowledge bases directed towards decision support in enterprise information systems. An approach to finding solutions in knowledge bases using document metadata was considered. Document metadata, as well as search queries, were represented as frames of design and search patterns respectively. Slots of those patterns correspond to the concepts' roles in the used ontology. The specified roles divide the concepts of ontology, document and query into clusters. Semantic networks for those clusters were defined in such a way that search queries of a document were represented as a set of semantic networks corresponding to the slots of design and search patterns. The relevancy of a document was estimated by a set of metrics which formalising proximity of semantic networks. Problem formulation of multi-criteria estimation of relevancy of documents in an enterprise ontological knowledge base and an adaptive solution method were presented in this paper.

Publications with keywords: semantic network, ontology, decision support
Publications with words: semantic network, ontology, decision support

References

1. Norenkov I.P. Intellektual'nye tekhnologii na baze ontologii [Intellectual technologies on the base of ontologies]. *Informatsionnye tekhnologii*, 2010, no. 1, pp. 17-23.

2. Tolcheev V.O. Metody vyyavleniya informatsionnykh priznakov v zadachakh klassifikatsii tekstovykh dokumentov [Methods of feature selection in text categorization tasks]. *Informatsionnye tekhnologii*, 2005, no. 8, pp. 14-21.

3. *The Dublin Core® Metadata Initiative*. Available at: http://dublincore.org/ , accessed 01.10.2013.

4. Karpenko A.P., Sokolov N.K. Otsenka slozhnosti semanticheskoy seti v obuchayushchey sisteme [Complexity estimation 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 01.10.2013.

5. Karpenko A.P., Sokolov N.K. Rasshirennaya semanticheskaya set' obuchayushchey sistemy i otsenka ee slozhnosti [Expanded semantic network of a tutoring system and its complexity measures]. *Nauka i obrazovanie MGTU im. N.E. Baumana* [Science and Education of the Bauman MSTU], 2008, no. 12. Available at: http://technomag.edu.ru/doc/111716.html , accessed 01.10.2013.

6. Galyamova E.V., Karpenko A.P., Sokolov N.K. Metodika kontrolya ponyatiynykh znaniy sub"ekta obucheniya v obuchayushchey sisteme [Technique of the control of conceptual knowledge of the subject of training in training 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 01.10.2013.

7. Karpenko A.P., Sokolov N.K. Mery slozhnosti semanticheskoy seti v obuchayushchey sisteme [Complexity Measures of Semantic Network of Learning System]. *Vestnik MGTU im. Baumana. Ser. Mashinostroenie.* [Herald of the Bauman MSTU. Ser. Instrument Engineering], 2009, no. 1 (74), pp. 50-66.

8. Galyamova E.V., Karpenko A.P., Sokolov N.K., Yagudaev G.G. Kontrol' ponyatiynykh znaniy sub"ekta obucheniya v obuchayushchey sisteme [Control of conceptual knowledge of the subject of training in training system]. *Vestnik MADI (GTU)*, 2009, no. 2(17), cpp 82-86.

9. Kogalovskiy M.R. *Perspektivnye tekhnologii informatsionnykh system* [Prospective technologies of information systems]. Moscow, DMK Press; Kompaniya AyTi Publ., 2003. 288 p.

10. Karpenko A.P. Otsenka relevantnosti dokumentov ontologicheskoy bazy znaniy [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 01.10.2013.

11. Lotov A.V., Pospelova I.I. *Mnogokriterial'nye zadachi priniatiia reshenii* [Multicriterion problems of decision making]. Moscow, MAKS Press, 2008. 197 p.

12. Zhang Q., Li H. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation*, 2007, vol. 11, no. 6, pp. 712-731. DOI: 10.1109/TEVC.2007.892759

13. Karpenko A.P., Mukhlisullina D.T., Ovchinnikov V.A. Neyrosetevaya approksimatsiya funktsii predpochteniy litsa, prinimayushchego resheniya, v zadache mnogokriterial'noy optimizatsii [Neural network approximation of decisions maker's utility function in multicriteria optimization problem]. *Informatsionnye tekhnologii*, 2010, no. 10, pp. 2-9.

14. Mukhlisullina D.T., Moor D.A., Karpenko A.P. Mnogokriterial'naya optimizatsiya na osnove nechetkoy approksimatsii funktsii predpochteniy litsa, prinimayushchego resheniya [Multi-criteria optimization based on fuzzy approximation of the preferences function of a decision maker]. *Nauka i obrazovanie MGTU im. N.E. Baumana* [Science and Education of the Bauman MSTU], 2010, no. 1. Available at: http://technomag.edu.ru/doc/135375.html , accessed 01.10.2013.

15. Karpenko A.P., Fedoruk V.G. Approksimatsiya funktsii predpochteniy litsa, prinimayushchego resheniya, v zadache mnogokriterial'noy optimizatsii. 3. Metody na osnove neyronnykh setey i nechetkoy logiki [Approximation of functions of the preferences of the decision maker in multicriteria optimization problem. 3. Methods based on neural networks and fuzzy logic]. *Nauka i obrazovanie MGTU im. N.E. Baumana* [Science and Education of the Bauman MSTU], 2008, no. 4. Available at: http://technomag.edu.ru/doc/86335.html , accessed 01.10.2013.

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