электронное научно-техническое издание
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 30569. Государственная регистрация №0420900025. ISSN 1994-0408

Оценка релевантности документов онтологической базы знаний
 # 09, сентябрь 2010
автор: Карпенко А. П.

УДК 519.6

МГТУ им. Н.Э. Баумана,
apkarpenko@mail.ru

Введение

Корпоративная база знаний представляет собой, как правило, совокупность разного рода слабоструктурированных документов, в которых с той или иной степенью подробности описаны прецеденты – 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 этих базах наиболее подходящих прецедентов и соответствующих им документов [1]. В работе рассматривается поиск решений по атрибутам документов, содержащимся в их метаданных, как альтернатива полнотекстовому поиску.

Классический атрибутивный поиск основывается на использовании в качестве метаданных документа преимущественно его регистрационных атрибутов [2]. В работе рассматривается иной подход к поиску решений в базах знаний прецедентов, когда метаданные формируются на основе онтологии соответствующей предметной области, заданной в виде семантической сети. При этом релевантность документа оценивается близостью в некоторой метрике семантической сети этого документа и семантической сети запроса.

В работе существенно используется «важность» концептов в семантической сети рассматриваемой онтологической базы знаний. Ряд мер этой важности предложен в нашей работе [3].

Важной составной частью предлагаемой методики оценки релевантности документа является построение семантической сети этого документа. В работе данная задача ставится, как задача огрубления графа семантической сети онтологии рассматриваемой предметной области [4,5]. Рассматриваются три метода решения задачи на основе насыщенных паросочетаний – методы, использующие случайные паросочетания, паросочетания из тяжелых ребер, а также паросочетания из тяжелых клик [4, 6,7].

В общей постановке о задаче поиска информации следует говорить в терминах модели поиска, которая включает в себя способ представления документов, способ представления поисковых запросов, вид критерия релевантности документов [8].

В данной работе документы в базе знаний, а также поисковые запросы к базе представляются в виде фреймов, которые называются паттерном проектирования и паттерном запроса соответсвенно. 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 работе предложено несколько мер релевантности ролевых кластеров документа, формализующих близость семантических сетей поискового образа документа и семантических сетей запроса. На основе указанных мер предложен алгоритм оценки релевантности документа запросу.

1. Построение семантической сети документа

Представим семантическую сеть $S(O)$ рассматриваемой онтологии O в виде взвешенного связного мультиграфа $G(O)$**.** **У**злы этого графа соответствуют концептам множества $C(O)=\{c_i, i\in[1:n^O]\}$, а ребра – четким бинарным отношениям между ними, каждое из которых принадлежит одному из типов $U_p, p\in[1:m^O]$.

Определены веса w_i^O, $i\in[1:n^O]$ узлов графа $G(O)$, формализующие «важность» соответствующих концептов в сети $S(O)$. Для каждого из ребер (c_i,c_j), $i,j\in[1:n^O]$, $i\neq j$ графа $G(O)$ полагается заданным также $(1\times m^O)$-вектор весов $\{v_{i,j,p}^O, p\in[1:m^O]\}$, где $v_{i,j,p}^O=0$, если концепты (c_i,c_j) не связаны между собой отношением типа U_p, и $v_{i,j,p}^O=v_p^O$ - в противном случае. Здесь v_p^O - априори заданный вес отношений типа U_p в онтологии O.

Некоторые методы определения весов концептов w_i^O и весов отношений v_p^O предложены в работе [3]. Для определения весов концептов может быть также использована их семантическая близость, полученная с помощью соответствующего словаря [9] или Википедии [10]. Веса концептов могут быть сформированы также на основе понятий центральности по близости и центральности по посредничеству [11].

Перейдем, например, с помощью аддитивной свертки

$$v_{i,j}^O=\sum_p \lambda_p v_{i,j,p}^O,\ p\in[1:m^O] \qquad (1)$$

от взвешенного мультиграфа $G(O)$ к взвешенному обыкновенному графу, в котором вес ребра (c_i,c_j) равен $v_{i,j}^O$. Сохраним за полученным графом прежнее обозначение. Здесь и далее λ_p, $p=1,2,...$ - положительный скалярный вещественный множитель, определяющий относительный вес компонентов аддитивной скалярной свертки вида (1).

Аналогично определим семантическую сеть $S(T) \subset S(O)$ документа T в виде связного взвешенного обыкновенного графа $G(T)$. Узлы этого графа соответствуют $n^T \le n^O$ концептам $C(T) \subset C(O)$ документа T, а ребра – связям между ними. Вес узла графа $G(T)$, соответствующего концепту $c_i \in C(T)$, обозначим w_i^T, а атрибуты его ребра (c_i, c_j) зададим парой $(l_{i,j}^T; v_{i,j}^T)$, где $l_{i,j}^T$ - «расстояние» между узлами c_i, c_j, а $v_{i,j}^T$ - вес ребра (c_i, c_j).

В терминах графа $G(T)$ задача построения семантической сети документа $S(T)$ сводится к решению двух следующих задач.

Задача 1 (задача определения топологии графа $G(T)$) - по каким правилам связывать узлы этого графа ребрами, т.е. устанавливать связи между концептами множества $C(T)$?

Задача 2 (задача определения весов узлов и атрибуты ребер графа $G(T)$) - исходя из каких соображений, назначать веса w_i^T узлов этого графа, а также атрибуты $l_{i,j}^T$, $v_{i,j}^T$ его ребер?

1.1. Определение топологии графа $G(T)$. В общей постанове эту задачу следует отнести к задаче огрубления графа [4].

Классические методы решения задачи огрубления графа основаны на итерационном стягивании смежных узлов графа G^α в узлы графа $G^{\alpha+1}$, где $\alpha = 0,1,2,...$ - номер итерации, $G^0 = G(O)$. В результате этого процесса ребро между двумя вершинами графа G^α удаляется и создается мультиузел графа $G^{\alpha+1}$, объединяющий оба стягиваемых узла. Задача огрубления графа $G(O)$ до графа $G(T)$ имеет ту специфику, что ни на одной из итераций указанного итерационного процесса в один узел не могут быть стянуты те узлы графа G^α, которые принадлежат графу $G(T)$.

Обычно задача огрубления графа решается в терминах паросочетаний. *Паросочетанием* в графе называется набор его ребер, в котором любые два

ребра не инцидентны общему узлу. Таким образом, граф $G^{\alpha+1}$ строится на основе графа G^{α} путем нахождения в графе G^{α} паросочетания и стягивания в мультиузел узлов, входящих в каждую из пар этого паросочетания. Непарные узлы графа G^{α} просто копируются в граф $G^{\alpha+1}$. Важно, что граф, огрубленный с использованием паросочетаний, сохраняет многие свойства исходного графа. Так, например, если граф G^{0} является планарным, то граф G^{α} также планарен [12].

В терминах паросочетаний специфика нашего случая состоит в том, что любая пара узлов каждого из паросочетаний не может включать в себя одновременно два узла графа $G(T)$.

С точки зрения повышения эффективности процесса огрубления графа, целесообразно использовать *насыщенные паросочетания* - паросочетания в которых хотя бы один узел любого ребра, не вошедшего в паросочетание, инцидентен ребру, вошедшему в паросочетание. Вообще говоря, с той же точки желательным является использование *максимальных паросочетаний* - насыщенных паросочетаний, которые имеют максимальное число ребер. Однако, вычислительная сложность формирования максимальных паросочетаний, в общем случае, значительно выше аналогичной вычислительной сложности для просто насыщенных паросочетаний. Поэтому обычно в вычислительной практике ограничиваются последними [7].

Утверждение 1. Оценка снизу количества итераций, необходимых для построения графа $G(T)$ с использованием насыщенных паросочетаний равна $\log\left(n^{O}/n^{T}\right)$.

Справедливость утверждения следует из того факта, что при использовании насыщенных паросочетаний число узлов графа $G^{\alpha+1}$ не может быть, очевидно, меньше половины числа узлов графа G^{α}.

Наиболее известны три следующих метода построения насыщенных паросочетаний: случайное паросочетание (*RM*); паросочетание из тяжелых ребер (*HEM*); паросочетание из тяжелых клик (*HCM*) [5].

Случайное паросочетание на итерации α строится по следующей схеме:

1) все узлы C^{α} текущего графа G^{α} объявляем немаркированными;
2) случайным образом выбираем немаркированный узел, еще не включенный в паросочетание - пусть это будет узле c_i^{α};
3) из числа немаркированных узлов, смежных узлу c_i^{α}, случайным образом выбираем узел (пусть это будет узел c_j^{α}), также еще не включенный в паросочетание;
4) если оба узла или один из узлов пары $c_i^{\alpha}, c_j^{\alpha}$ не принадлежат графу $G(T)$, то включаем ребро $(c_i^{\alpha}, c_j^{\alpha})$ в паросочетание, и узлы $c_i^{\alpha}, c_j^{\alpha}$ маркируем;
5) если ни одного немаркированного узла, смежного узлу c_i^{α}, не существует, то узел c_i^{α} маркируем и оставляем свободным (чтобы затем перенести его в граф $G^{\alpha+1}$);
6) если в графе G^{α} имеются еще немаркированные узлы, то переходим к шагу 2.

Данную схему иллюстрирует рисунок 1, на котором слева показан граф $G^{\alpha-1}$ и сформированное на его основе паросочетание, а справа – граф G^{α}.

*Паросочетание из тяжелых ребер***.** Схема построения этого паросочетания отличается от рассмотренной выше схемы шагом 3, который в данном случае формулируется следующим образом. Из числа немаркированных узлов, смежных узлу c_i^{α}, выбираем такой узел c_j^{α}, еще не включенный в паросочетание, что вес ребра $(c_i^{\alpha}, c_j^{\alpha})$ является максимальным среди весов всех возможных ребер, связанных с узлом c_i^{α}.

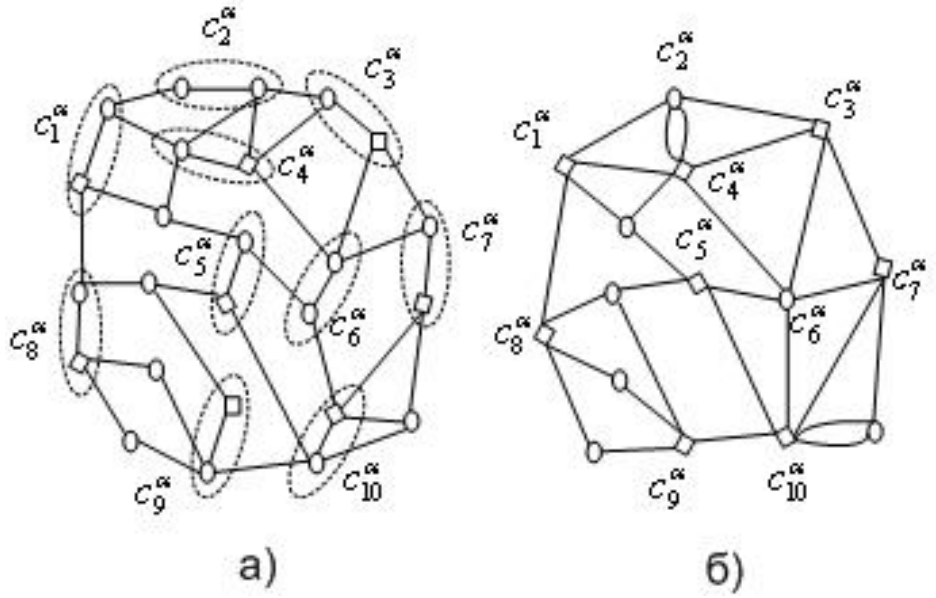

Рисунок 1 – К методу случайных паросочетаний: квадратиками показаны узлы графа $G(T)$: а) граф $G^{\alpha-1}$; б) граф G^{α}

Паросочетание из тяжелых клик. В данном случае также меняется только шаг 3 рассмотренной схемы формирования случайного паросочетания: из числа немаркированных узлов, смежных узлу c_i^{α}, выбираем такой узел c_j^{α}, еще не включенный в паросочетание, что реберная плотность мультиузла, который получается стягиванием узлов $c_i^{\alpha}, c_j^{\alpha}$, является максимально возможной по сравнению со всеми иными вариантами выбора узла c_j^{α}.

Итерации во всех рассмотренных методах формирования паросочетания заканчиваются, когда в результате данной итерации не удалось выделить ни одной пары узлов. Другими словами, итерации заканчиваются, если в текущем графе G^{α} содержатся только узлы графа $G(T)$.

Отметим следующее обстоятельство. В силу наличия элемента случайности при формировании паросочетаний, различные итерационные процессы порождают, вообще говоря, графы $G(T)$, имеющие различную топологию. Таким образом, возникает задача получения в некотором смысле наилучшего графа $G(T)$. При этом в качестве максимизируемого критерия

оптимальности графа можно использовать, например, его реберную плотность (коэффициент кластеризации) [3,11].

1.2. Определение весов узлов и ребер графа $G(T)$. Выделим два случая:

1) рассматриваемая пара узлов паросочетания включает в себя узел, принадлежащий графу $G(T)$ (например, пара c_1^α на рисунке 1);
2) пара узлов содержит только узлы, не принадлежащие графу $G(T)$ (например, пара c_2^α на том же рисунке).

Случай 1. Пусть рассматриваемая пара включает в себя узел (или мультиузел) $c_i^\alpha \in C(T)$ и узел (или мультиузел) $c_j^\alpha \notin C(T)$, веса которых равны w_i^α, w_j^α соответственно, а атрибуты ребра (c_i^α, c_j^α) определяется парой $(l_{i,j}^\alpha; v_{i,j}^\alpha)$. Отметим, что во веденных обозначениях надиндекс α указывает на то, что в процессе огрубления графа $G(O)$ веса его узлов и ребер, вообще говоря, изменяются. Здесь и далее в данном разделе для простоты записи индекс T в обозначениях опущен.

Будем полагать, что в процессе стягивания узлов c_i^α, c_j^α узел c_j^α стягивается к c_i^α, так что в результате получается мультиузел $c_i^{\alpha+1} \in C(T)$, вес которого равен $w_i^{\alpha+1}$. Условно результат данной процедуры будем записывать в виде $c_i^{\alpha+1} = c_i^\alpha \oplus c_j^\alpha$.

Логично исходить из того, что вес $w_i^{\alpha+1} = w_i^{\alpha+1}(w_i^\alpha, w_j^\alpha, l_{i,j}^\alpha, v_{i,j}^\alpha)$ является некоторой положительной возрастающей функций своих аргументов w_i^α, w_j^α, $v_{i,j}^\alpha$ и такой же убывающей функций аргумента $l_{i,j}^\alpha$. В простейшем случае в качестве такой функции может быть использована функция вида

$$w_i^{\alpha+1}(w_i^\alpha, w_j^\alpha, l_{i,j}^\alpha, v_{i,j}^\alpha) = \lambda_1 w_i^\alpha + \lambda_2 w_j^\alpha \frac{v_{i,j}^\alpha}{l_{i,j}^\alpha}. \qquad (2)$$

В результате стягивания узла c_j^α к узлу c_i^α атрибуты ребер, инцидентных узлу c_i^α, не меняются, а значения атрибутов ребер, инцидентных узлу c_j^α, должны быть по некоторому правилу изменены. Рассмотрим одно из таких ребер (c_j^α, c_p^α), атрибуты которого равны $(l_{j,p}^\alpha; v_{i,p}^\alpha)$. Это ребро заменяется на ребро $(c_i^{\alpha+1} = c_i^\alpha \oplus c_j^\beta, c_p^{\alpha+1} = c_p^\alpha)$, которому соответствуют атрибуты $(l_{i,p}^{\alpha+1}; v_{i,p}^{\alpha+1})$.

Естественно положить, что длина ребра $(c_i^{\alpha+1}, c_p^{\alpha+1})$ равна

$$l_{i,p}^{\alpha+1} = l_{i,j}^\alpha + l_{j,p}^\alpha,$$

т.е. на длину ребра (c_i^α, c_j^α) превышает длину ребра (c_j^α, c_p^α). Логично также принять, что вес ребра $v_{i,p}^{\alpha+1} = v_{i,p}^{\alpha+1}(v_{i,j}^\alpha, v_{j,p}^\alpha)$ является некоторой положительной возрастающей функцией своих аргументов. В простейшем случае можно положить

$$v_{i,p}^{\alpha+1} = \lambda_1 v_{i,j}^\alpha + \lambda_2 v_{j,p}^\alpha. \quad (3)$$

Схему рассмотренного алгоритма иллюстрирует рисунок 2. Здесь принято, что

$$w_i^{\alpha+1}(w_i^\alpha, w_j^\alpha, l_{i,j}^\alpha, v_{i,j}^\alpha) = w_i^\alpha + w_j^\alpha \frac{v_{i,j}^\alpha}{l_{i,j}^\alpha}, \quad (4)$$

$$v_{i,p}^{\alpha+1} = v_{i,j}^\alpha + v_{j,p}^\alpha. \quad (5)$$

Случай 2. Положим, что оба узла рассматриваемой пары (c_i^α, c_j^α) не принадлежат графу $G(T)$, т.е. когда имеет место ситуация $c_i^\alpha, c_j^\alpha \notin C(T)$. Как и ранее, положим, что веса указанных узлов равны w_i^α, w_j^α, а атрибуты ребра (c_i^α, c_j^α) определяется парой $(l_{i,j}^\alpha; v_{i,j}^\alpha)$.

В этом случае также можно считать, что один из узлов (пусть это будет узел c_j^α) стягивается к другому узлу (c_i^α), так что в результате получается мультиузел $c_i^{\alpha+1} = c_i^\alpha \oplus c_j^\alpha$, $c_i^{\alpha+1} \notin C(T)$ Как и в предыдущем случае, положим,

что вес этого узла равен $w_i^{\alpha+1}$ и представляет собой, например, функцию вида (2).

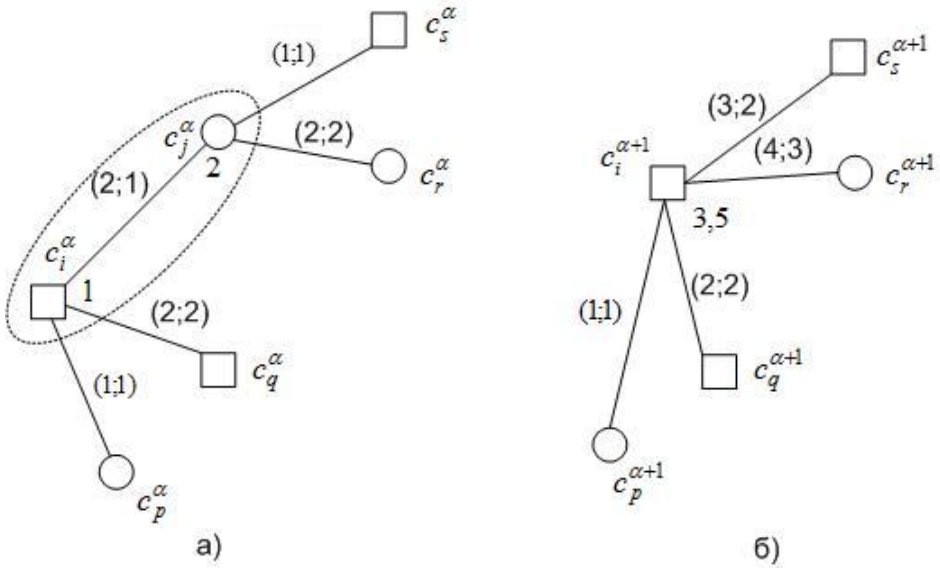

Рисунок 2 – К стягиванию узла (мультиузла) $c_i^\alpha \in C(T)$ и узла (мультиузла) $c_j^\alpha \notin C(T)$: квадратиками показаны узлы графа $G(T)$

В отличие от предыдущего случая, здесь логично положить, что в результате стягивания узла c_j^α к узлу c_i^α меняются значения атрибутов всех ребер, инцидентных как узлу c_i^α, так и узлу c_j^α. Рассмотрим ребра (c_i^α, c_p^α), (c_j^α, c_q^α), атрибуты которых равны $(l_{i,p}^\alpha; v_{i,p}^\alpha)$, $(l_{j,q}^\alpha; v_{j,q}^\alpha)$ соответственно. Эти ребра заменяются на ребра $(c_i^{\alpha+1}, c_p^{\alpha+1})$, $(c_i^{\alpha+1}, c_q^{\alpha+1})$, которым соответствуют атрибуты $(l_{i,p}^{\alpha+1}; v_{i,p}^{\alpha+1})$, $(l_{i,q}^{\alpha+1}; v_{i,q}^{\alpha+1})$, где

$$l_{i,p}^{\alpha+1} = 0{,}5 l_{i,j}^\alpha + l_{i,p}^\alpha, \qquad l_{i,q}^{\alpha+1} = 0{,}5 l_{i,j}^\alpha + l_{j,q}^\alpha, \tag{6}$$

а веса $v_{i,p}^{\alpha+1}$, $v_{i,q}^{\alpha+1}$ определяются, например, по формуле вида (2).

Отметим, что принятые соглашения могут приводить к нецелым значениям расстояний между узлами графа $G^{\alpha+1}$, даже если все расстояния между узлами графа G^α являются целыми.

Схему рассмотренного алгоритма иллюстрирует рисунок 3. Здесь принято, что вес узла $c_i^{\alpha+1}$ определяется по формуле (4), веса ребер графа $G^{\alpha+1}$ - по формуле (5), а расстояние между узлами этого графа – по формуле (6).

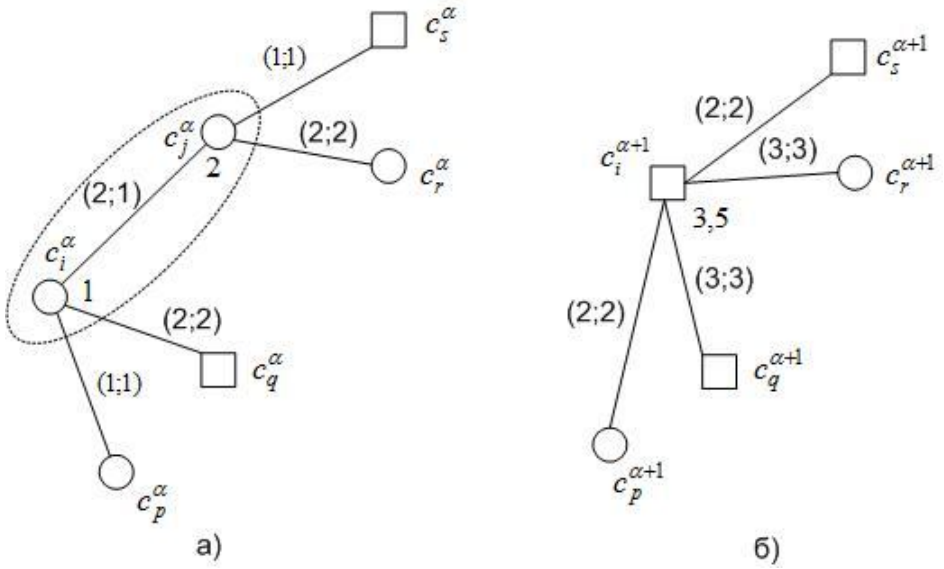

Рисунок 3 – К стягиванию узлов (мультиузлов) $c_i^{\alpha}, c_j^{\alpha} \notin C(T)$

В результате итерации α в графе $G^{\alpha+1}$ могут появиться кратные ребра (см., например, узлы c_2^{α}, c_4^{α} на рисунке 1а). Прежде чем переходить к основному циклу итерации $(\alpha+1)$ эти ребра следует объединить. Возникает вопрос, как вычислить значения атрибутов полученного ребра?

Положим, что двумя ребрами связаны узлы c_i^{α}, c_j^{α}, и атрибуты этих ребер равны $(l_{1,i,j}^{\alpha}, v_{1,i,j}^{\alpha})$, $(l_{2,i,j}^{\alpha}, v_{2,i,j}^{\alpha})$. В качестве расстояния между этими узлами $l_{i,j}^{\alpha+1}$ примем минимальное из расстояний $l_{1,i,j}^{\alpha}$, $l_{2,i,j}^{\alpha}$:

$$l_{i,j}^{\alpha+1} = \min(l_{1,i,j}^{\alpha}, l_{2,i,j}^{\alpha}).$$

В качестве веса $v_{i,j}^{\alpha+1}$ логично принять сумму весов указанных ребер:

$$v_{i,j}^{\alpha+1} = v_{1,i,j}^{\alpha} + v_{2,i,j}^{\alpha}.$$

Таким образом, после завершения итераций оказываются полностью определенными топология графа $G(T)$, а также веса его узлов w_i и значения атрибутов его ребер $(l_{i,j}, v_{i,j})$; $i, j \in [1:n^T]$, $i \neq j$.

Вернемся к использованию в обозначениях весов узлов и атрибутов ребер графа $G(T)$ индекса T.

Исключим из числа атрибутов ребер графа $G(T)$ расстояния $l_{i,j}^T$ и модифицируем веса $v_{i,j}^T$ ребер этого графа: положим, что «новый» вес ребра (c_i, c_j) равен $v_{i,j}^T = v(l_{i,j}^T, v_{i,j}^T)$, где $v(l_{i,j}^T, v_{i,j}^T)$ - некоторая положительная убывающая функция расстояния $l_{i,j}^T$ и такая же возрастающая функция «старого» веса $v_{i,j}^T$. Например, можно принять

$$v_{i,j}^T = \lambda_1 \frac{v_{i,j}^T}{l_{i,j}^T}. \qquad (7)$$

При необходимости, можно нормировать веса узлов и ребер полученного графа $G(T)$, например, следующим образом:

$$w_i^T = \frac{w_i^T}{w_{\max}^T};\ v_{i,j}^T = \frac{v_{i,j}^T}{v_{\max}^T};\ i, j \in [1:n^T];\ i \neq j.$$

Здесь $w_{\max} = \max_i w_i^T$, $v_{\max}^T = \max_{i,j} v_{i,j}^T$ - максимальный вес узла и ребра в графе $G(T)$ соответственно.

2. Ролевая кластеризация семантических сетей онтологии и документа

Положим, что выделено k ролей ω_i, $i \in [1:k]$ концептов. Роли ω_i разбивают все множество концептов $C(O)$ на k непересекающихся «ролевых» кластеров D_i^O, среди которых могут быть и пустые кластеры. Множество концептов, принадлежащих кластеру D_i^O, обозначим C_i^O, так что

$$C(O) = \bigcap_{i=1}^{k} C_i^O .$$

Число концептов в кластере D_i^O (или, что то же самое, во множестве C_i^O) обозначим n_i^O. Очевидно, что

$$\sum_{i=1}^{k} n_i^O = n^O .$$

Аналогично, роли ω_i разбивают множество концептов C^T документа T на k ролевых кластеров D_i^T, концепты которых образуют множества C_i^T с числом концептов в них, равным n_i^T:

$$C^T = \bigcap_{i=1}^{k} C_i^T ; \qquad \sum_{i=1}^{k} n_i^T = n^T .$$

Кластерам D_i^O, D_i^T поставим в соответствие их семантические сети S_i^O, S_i^T и графы G_i^O, G_i^T; $i \in [1:k]$. Обозначим $w_{i,p}^O$ - вес узла $c_{i,p}$ графа G_i^O, $v_{i,p,q}^O$ - вес ребра графа G_i^O, связывающего его узлы $c_{i,p}, c_{i,q}$. Здесь $p,q \in [1:n_j^O], p \neq q$. Аналогичные обозначения $w_{i,p}^T$, $v_{i,p,q}^T$ введем для графа G_i^T.

Графы G_i^O, G_i^T, $i \in [1:k]$ ролевых кластеров D_i^O, D_i^T могут быть построены по схеме, рассмотренной в п.1. При этом графы G_i^O строятся на основе графа $G(O)$, а графы G_i^T - на основе графа G^T.

Отметим, что оценить качество рассмотренной ролевой кластеризации можно, например, с помощью величины, которая называется модулярность (modularity) графа [13].

3. Поисковые образы документа и запроса

Пусть в семантической сети документа T выделены ролевые кластеры D_i^T и тем или иным образом построены семантические сети этих кластеров S_i^T, а также соответствующие им графы G_i^T; $i \in [1:k]$. Положим, что паттерн

проектирования $A(T) = A$ документа T имеет k слотов $A_i(T) = A_i$ и слот A_i соответствует роли ω_i.

Поисковый образ рассматриваемого документа T будем представлять в виде k семантических сетей S_i^T, формализованных в виде графов G_i^T; $i \in [1:k]$ - рисунок 4.

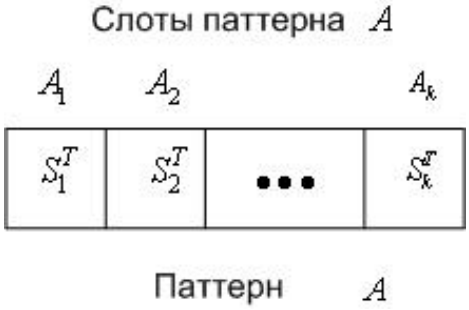

Рисунок 4 – Поисковый образ документа T_i

Не ограничивая общности рассуждений, положим, что поисковый образ запроса Q формируется паттерном $B(Q) = \{B_i(Q), i \in [1:k]\}$, который также имеет k слотов $B_i(Q) = B_i$.

Введем следующие обозначения:

C^Q - множество концептов запроса Q;

n^Q - число концептов во множестве C^Q;

D_i^Q - ролевые кластеры множества C^Q, $i \in [1:k]$;

C_i^Q - множество концептов кластера D_i^Q,

$$C^Q = \bigcap_{j=1}^{k} C_i^Q;$$

n_i^Q - число концептов в кластере D_i^Q,

$$\sum_{i=1}^{k} n_i^Q = n^Q;$$

S_i^Q - семантическая сеть кластера D_i^Q;

G_i^Q - граф семантической сети S_i^Q;

$w_{i,p}^{Q}$ - вес узла $c_{i,p}^{Q}$ графа G_i^{Q};

$v_{i,p,q}^{Q}$ - вес ребра $(c_{i,p}^{Q}, c_{i,q}^{Q})$ графа G_i^{Q}.

Здесь $p,q \in [1:n_i^{Q}]$, $p \neq q$.

Таким образом, поисковый образ запроса Q представляет собой k семантических сетей S_i^{Q}, формализованных в виде графов G_i^{Q}; $i \in [1:k]$ - рисунок 5.

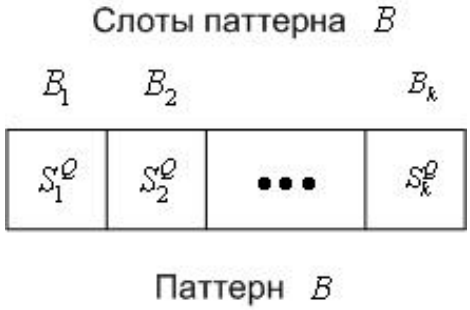

Рисунок 5 – Поисковый образ запроса Q

Графы G_i^{Q}, $i \in [1:k]$ ролевых кластеров D_i^{Q} также могут быть построены по схеме, рассмотренной в п.1 на основе графа $G(O)$.

4. Релевантность ролевого кластера документа

Можно предложить несколько мер релевантности ролевых кластеров документов, формализующих близость семантических сетей S_i^{T} поискового образа документа T и семантических сетей S_i^{Q} запроса Q или, что то же самое, мер близости соответствующих графов G_i^{T}, G_i^{Q}. Обозначим эти меры $r(S_i^{T}, S_i^{Q}) = r_i$.

Определим прежде меру близости концептов множества C_i^{O}

$$l(c_{i,p}, c_{i,q}) = l_{i,p,q} = \min(v_{i,p,\alpha}^{O} + v_{i,\alpha,\beta}^{O} + \ldots + v_{i,\zeta,q}^{O}),$$

где минимум берется по всем возможным цепям $c_{i,p}, c_{i,\alpha}, c_{i,\beta}, \ldots, c_{i,\zeta}, c_{i,q}$, в которых все концепты принадлежат множеству C_i^{O}.

Во множестве C_i^T найдем для концепта $c_{i,p}$, такого что $c_{i,p} \in C_i^Q$, $c_{i,p} \notin C_i^T$, концепт $c_{i,q}$, расстояние которого до концепта $c_{i,p}$ равно $l_{i,p,q} = l_{i,p,q}^1$. Включим полученный концепт $c_{i,q}$ во множество $D_{1,i}^Q$. Повторим процедуру для всех концептов множества C_i^Q, которые не принадлежат множеству C_i^T. Полученный в результате кластер $D_{1,i}^Q$ представляет собой совокупность концептов множества C_i^T, не принадлежащих множеству C_i^Q, но находящихся ближе всего (в смысле меры l) к этому множеству. Положим, что мощность кластера $D_{1,i}^Q$ равна $n_{1,i}^Q$.

Аналогично, для каждого концепта $c_{i,p}$, такого что $c_{i,p} \in C_i^Q$, $c_{i,p} \notin C_i^T$, найдем во множестве $C_i^T \setminus D_{1,i}^Q$ концепт $c_{i,q}$, расстояние которого до концепта $c_{i,p}$ равно $l_{i,p,q} = l_{i,p,q}^2$ и включим все полученные концепты во множество $D_{2,i}^Q$. Кластер $D_{2,i}^Q$ представляет собой совокупность концептов множества C_i^T, не принадлежащих множествам C_i^Q, $D_{1,i}^Q$, но находящихся ближе всего (в смысле той же меры l) к кластеру $D_{1,i}^Q$. Мощность кластера $D_{2,i}^Q$ равна $n_{2,i}^Q$. И.т.д.

Взаимосвязи кластеров D_i^Q, $D_{1,i}^Q$, $D_{2,i}^Q$,…, D_i^T, D_i^O иллюстрирует рисунок 6.

Для каждого из концептов $c_{i,p} \in C_i^Q$, $c_{i,p} \notin C_i^T$ и концептов $c_{i,q} \in D_{t,i}^Q$ определим функцию $f_{i,p,q}^t(w_{i,q}^T, l_{i,p,q}^t)$, которая является положительной возрастающей функцией относительно первого аргумента и такой же убывающей функцией относительно второго аргумента; $t = 1,2,...$ Примером такой функции может служить функция

$$f_{i,p,q}^t(w_{i,q}^T, l_{i,p,q}^t) = f_{i,p,q}^t = \lambda_1 \frac{w_{i,q}^T}{l_{i,p,q}^t}. \quad (8)$$

Функция $f_{i,p,q}^{t}$ формализует уменьшение весов концептов из кластеров $D_{t,i}^{Q}$ по мере «удаления» их от кластера D_i^Q.

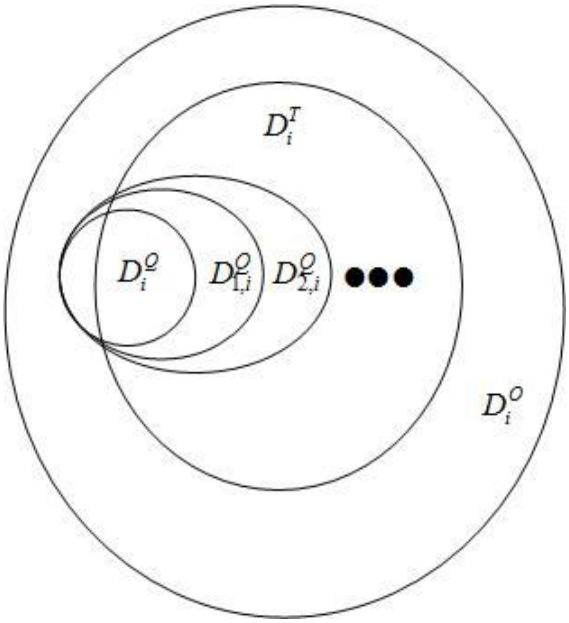

Рисунок 6 – К взаимосвязям кластеров D_i^Q, $D_{1,i}^Q$, $D_{2,i}^Q$,…, D_i^T, D_i^O

4.1. Меры, не учитывающие веса концептов и связей между ними.

Мера на основе коэффициента Дайса, используемого при сравнении текстовых документов [14]

$$r_{i,1} = \frac{2n(G_i^Q \bigcap G_i^T)}{n(G_i^Q) + n(G_i^T)} = \frac{2n(G_i^Q \bigcap G_i^T)}{n_i^Q + n_i^T} \in [0,1], \qquad (9)$$

где $n(G_i^Q) = n_i^Q$, $n(G_i^T) = n_i^T$ - числа узлов графов G_i^Q, G_i^T, соответственно; $n(G_i^Q \bigcap G_i^T)$ - числа узлов, содержащихся как в графе G_i^Q, так и в графе G_i^T.

Мера (9), по сути, представляет собой относительное число концептов кластера D_i^Q, содержащихся в кластере D_i^T, и в работе [14] называется мерой концептуальной близостью графов G_i^Q, G_i^T.

Здесь и далее полагается, что $i \in [1:k]$.

Мера на основе относительной близости графов G_i^Q, G_i^T [14]

$$r_{i,2} = \frac{2m(G_i^Q \bigcap G_i^T)}{m(G_i^Q) + m(G_i^T)} = \frac{2m(G_i^Q \bigcap G_i^T)}{m_i^Q + m_i^T} \in [0,1], \quad (10)$$

где $m(G_i^Q) = m_i^Q$, $m(G_i^T) = m_i^T$ - числа ребер, содержащихся в графах G_i^Q, G_i^T, соответственно; $n(G_i^Q \bigcap G_i^T)$ - число ребер, содержащихся как в графе G_i^Q, так и в графе G_i^T.

Известно, что меры вида (9), (10) сильно зависят от размеров графов [14]. Поэтому целесообразно использовать их следующие модификации, учитывающие размеры графов G_i^Q, G_i^T.

Модифицированная мера на основе меры $r_{i,1}$

$$r_{i,3} = \frac{2n(G_i^Q \bigcap G_i^T)e_1}{n_i^Q + n_i^T}, \quad (11)$$

где

$$e_1 = \begin{cases} \lambda_1 \dfrac{n_i^Q}{n_i^T}, & n_i^Q \geq n_i^T, \\ \lambda_1 \dfrac{n_i^T}{n_i^Q}, & n_i^T > n_i^Q. \end{cases} \quad (12)$$

Модифицированная мера на основе меры $r_{i,2}$

$$r_{i,4} = \frac{2m(G_i^Q \bigcap G_i^T)e_2}{m_i^Q + m_i^T}, \quad (13)$$

где величина e_2 определяется по формуле вида (12).

Очевидно, что при $e_1 = 1$ меры (12), (13) совпадают с мерами $r_{i,1}$, $r_{i,2}$ соответственно, так что последние меры являются частным случаем мер (9), (10).

Мера, являющаяся расширением меры $r_{i,3}$

$$r_{i,5} = r_{i,3} + \lambda_1 \frac{n_{1,i}^Q}{n_i^Q + n_{1,i}^Q} + \lambda_2 \frac{n_{2,i}^Q}{n_i^Q + n_{1,i}^Q + n_{2,i}^Q} + \ldots. \quad (14)$$

Мера имеет смысл относительного числа узлов графа G_i^Q, содержащихся в графе G_i^T, и графах $G_{1,i}^Q$, $G_{2,i}^Q$,….

Мера, являющаяся расширением меры $r_{i,4}$ и аналогичная мере (14)

$$r_{i,6} = r_{i,4} + \lambda_1 \frac{m_{1,i}^Q}{m_i^Q + m_{1,i}^Q} + \lambda_2 \frac{m_{2,i}^Q}{m_i^Q + m_{1,i}^Q + m_{2,i}^Q} + ... \qquad (15)$$

Здесь $m_{t,i}^Q$ - число ребер, содержащихся в графе $G_{t,i}^Q$; $t = 1,2,...$

Отметим, что, очевидно, меры (14), (15) являются частными случаями мер (11), (13).

На основе мер (9) – (11), (13) – (15) легко сконструировать меры, которые учитывают только «сильные» узлы и ребра в графах G_i^Q, $G_{1,i}^Q$, $G_{2,i}^Q$,…., т.е. узлы и ребра, веса которых превышают некоторые заданные величины [15].

Аддитивная свертка мер $r_{i,5}$, $r_{i,6}$

$$r_{i,7} = \lambda_1 r_{i,5} + \lambda_2 r_{i,6}, \qquad (16)$$

включающая в себя все рассмотренные выше меры.

4.2. Меры, учитывающие веса концептов и связей между ними.

Взвешенная мера на основе меры $r_{i,1}$

$$r_{i,8} = \frac{2\sum w_{i,\alpha}^Q}{\sum w_{i,\beta}^Q + \sum w_{i,\gamma}^T} \in [0,1], \qquad (17)$$

где индекс α пробегает номера узлов, принадлежащих пересечению графов $G_i^Q \bigcap G_i^T$, что условно будем записывать в виде $\alpha \in [1 : n(G_i^Q \bigcap G_i^T)]$; индексы β, γ пробегают номера узлов $[1 : n_i^Q]$, $[1 : n_i^T]$ соответственно.

Очевидно, что мера (17) эквивалентна мере (9), если принять следующие соглашения: $w_{i,\alpha}^Q = 1$ при $\alpha \in [1 : n(G_i^Q \bigcap G_i^T)]$; $w_{i,\alpha}^Q = 0$ - в противном случае; $w_{i,\beta}^T = 1$, $w_{i,\gamma}^T = 1$. Таким образом, меру (9) можно считать частным случаем меры (17).

Взвешенная мера на основе меры $r_{i,2}$

$$r_{i,9} = \frac{2\sum v^{Q}_{i,\alpha,\beta}}{\sum v^{Q}_{i,\gamma,\delta} + \sum v^{T}_{i,\varphi,\chi}} \in [0,1], \quad (18)$$

где, аналогично (17), $\alpha,\beta \in [1 : n(G_i^Q \bigcap G_i^T)]$, $\gamma,\delta \in [1 : n_i^Q]$, $\varphi,\chi \in [1 : n_i^T]$; $\alpha \neq \beta$, $\gamma \neq \delta$, $\varphi \neq \chi$. Легко видеть, что мера (18) является частным случаем меры (10).

Модифицированная мера на основе меры $r_{i,8}$

$$r_{i,10} = \frac{2\sum w^{Q}_{i,\alpha} a}{\sum w^{Q}_{i,\beta} + \sum w^{T}_{i,\gamma}}. \quad (19)$$

Модифицированная мера на основе меры $r_{i,9}$

$$r_{i,11} = \frac{2\sum v^{Q}_{i,\alpha,\beta} b}{\sum v^{Q}_{i,\gamma,\delta} + \sum v^{T}_{i,\varphi,\chi}}. \quad (20)$$

Мера, являющаяся расширением меры $r_{i,10}$

$$r_{i,12} = r_{i,10} + \lambda_1 \frac{\sum w^{Q}_{1,i,\beta}}{\sum w^{Q}_{i,\alpha} + \sum w^{Q}_{1,i,\beta}} + \lambda_2 \frac{\sum w^{Q}_{2,i,\gamma}}{\sum w^{Q}_{i,\alpha} + \sum w^{Q}_{1,i,\beta} + \sum w^{Q}_{2,i,\gamma}} + ..., \quad (21)$$

где $\alpha \in [1 : n(G_i^Q \bigcap G_i^T)]$; индекс β пробегает номера узлов, принадлежащих графу $G^Q_{1,i}$, что условно будем записывать в виде $\beta \in [1 : n(G^Q_{1,i})]$; аналогично $\gamma \in [1 : n(G^Q_{2,i})]$.

Мера, являющаяся расширением меры $r_{i,11}$

$$r_{i,13} = r_{i,11} + \lambda_1 \frac{\sum v^{Q}_{1,i,\gamma,\delta}}{\sum v^{Q}_{i,\alpha,\beta} + \sum v^{Q}_{1,i,\gamma,\delta}} + \lambda_2 \frac{\sum v^{Q}_{1,i,\varphi,\chi}}{\sum v^{Q}_{i,\alpha,\beta} + \sum v^{Q}_{1,i,\gamma,\delta} + \sum v^{Q}_{2,i,\varphi.\chi}} + ..., \quad (22)$$

аналогичная мере (21). Здесь $\alpha,\beta \in [1 : n(G_i^Q \bigcap G_i^T)]$, $\gamma,\delta \in [1 : n(G^Q_{1,i})]$, $\varphi,\chi \in [1 : n(G^Q_{2,i})]$; $\alpha \neq \beta$, $\gamma \neq \delta$, $\varphi \neq \chi$.

Аддитивная свертка мер $r_{i,12}$, $r_{i,13}$

$$r_{i,14} = \lambda_1 r_{i,12} + \lambda_2 r_{i,13}, \quad (23)$$

включающая в себя все рассмотренные выше меры (17) – (22).

Модифицированная мера на основе меры $r_{i,10}$

$$r_{i,15} = r_{i,10} + \lambda_1 \frac{\sum f^1_{i,\beta,\gamma}}{\sum w^Q_{i,\alpha} + \sum f^1_{i,\beta,\gamma}} + \lambda_2 \frac{\sum f^2_{i,\varphi,\chi}}{\sum w^Q_{i,\alpha} + \sum f^1_{i,\beta,\gamma} + \sum f^2_{i,\varphi,\chi}} + ..., \quad (24)$$

где $\alpha \in [1:n(G^Q_i \bigcap G^T_i)]$, $\beta,\gamma \in [1:n(G^Q_{1,i})]$, $\varphi,\chi \in [1:n(G^Q_{2,i})]$; $\alpha \neq \beta$, $\gamma \neq \delta$, $\varphi \neq \chi$.

Меры (17) – (24) также легко модифицировать, учитывая только «сильные» узлы и ребра в графах G^Q_i, $G^Q_{1,i}$, $G^Q_{2,i}$,…. [15]. Значительное число мер релевантности ролевого кластера документа может быть построено на основе мер семантической близости в сетях документов [16].

5. Оценка релевантности документа

Пусть поисковый образ документа T представлен паттерном проектирования $A = \{A_i, i \in [1:k]\}$, слотам которого A_i, $i \in [1:k]$ соответствуют семантические сети S^T_i, формализованные в виде графов G^T_i - рисунок 4. Пусть, аналогично, поисковый образ запроса Q сформирован в виде паттерна $B = \{B_i, i \in [1:k]\}$, который представляет собой совокупность k семантических сетей S^Q_i, формализованных в виде графов G^Q_i - рисунок 5.

Обозначим $R(T,Q) = R(r^Q_1, r^Q_2, ..., r^Q_k)$ релевантность документа T запросу Q, где $R(r^Q_1, r^Q_2, ..., r^Q_k)$ - некоторая неотрицательная вещественно значная возрастающая функция всех своих аргументов, например,

$$R(r^Q_1, r^Q_2, ..., r^Q_k) = \sum_{i=1}^{k} \lambda_i r^Q_i . \quad (25)$$

Нормировать величину $R(r^Q_1, r^Q_2, ..., r^Q_k)$ можно, отнеся ее к сумме релевантностей всех рассматриваемых документов базы знаний.

Общая схема предлагаемой методики оценки релевантности документа T имеет вид, представленный на рисунке 7.

Определение релевантности (25) можно расширить путем учета априорной «значимости» документа T, которую можно построить, например, на основе мер $\mu(S^O_i, S^T_i) = \mu^T_i$ близости семантических сетей S^O_i онтологии и

семантических сетей S_i^T документа T или, что то же самое, мер близости соответствующих графов G_i^O, G_i^T; $i \in [1:k]$. Так в качестве меры μ^T значимости документа T можно использовать подходящим образом нормированную взвешенную сумму мер $\mu_1^T, \mu_2^T, ..., \mu_k^T$:

$$\mu^T = \sum_{i=1}^{k} \lambda_i \mu_i^T . \qquad (26)$$

Рисунок 7 - Схема оценки релевантности документа

С учетом меры μ^T формула (25) модифицируется следующим образом:

$$R(r_1^Q, r_2^Q, ..., r_k^Q, \mu^T) = \mu^T \sum_{i=1}^{k} \lambda_i r_i^Q . \qquad (27)$$

Отметим, что формулы (25), (27) не учитывают эффективность решений, которые содержатся в документе T. На основе опыта эксплуатации рассматриваемой базы знаний эта эффективность может быть оценена лицом, принимающим решения и сохранена в базе знаний.

Заключение

Предложенная в работе методика оценки релевантности документов обладает высокой вычислительной сложностью. Подавляющая часть требуемых вычислительных затрат обусловлена выполнением следующих работ.

Во-первых, для каждого из документов T базы знаний методика требует построения соответствующей семантической сети $S(T)$, а также построения семантической сети S_i^T, $i \in [1:k]$ каждого из слотов поискового образа документа (паттерна проектирования). Если онтология предметной области фиксирована, то эта работа выполняется лишь однажды, при помещении документа в базу знаний.

Во-вторых, методика требует построения аналогичных семантических сетей S_i^O онтологии рассматриваемой предметной области. Опять же, если онтология предметной области фиксирована, то эта работа выполняется лишь однократно.

В-третьих, в соответствии с методикой для каждого из запросов Q также требуется формирование семантических сетей S_i^Q. Данная работа должна выполняться системой управления базой знаний при обработке каждого из запросов.

В работе широко используется аддитивная скалярная свертка (см., например, формулы (1), (2), (3), (7) и т.д.). Очевидно, что наряду с аддитивными свертками могут быть использованы и иные, например, мультипликативные свертки или их комбинация [17].

Основная задача работы – задача определения релевантности документа – является, по сути, задачей многокритериальной (точнее - k-критериальной)

оптимизации – см. формулы (25), (27). Использованный при решении этой задачи метод аддитивной скалярной свертки является простейшим и далеко не всегда эффективным методом решения многокритериальных задач. Поэтому представляет интерес исследование целесообразности использования других, более «тонких» методов решения указанной многокритериальной задачи [17].

Широкое использование сверток приводит к тому, что методика содержит большое число свободных параметров (см. формулы (1), (2), (3), (7) и т.д.). Имеется немного содержательных оснований для априорного выбора значений этих параметров. Поэтому представляется перспективным ставить задачу определения их значений, как задачу метаоптимизации [18]. Отметим, что при этом в базе знаний требуется хранить оценки успешности поиска, сформированные лицом, принимающим решения.

Одной из проблем, которая возникает при использовании рассмотренного подхода к определению релевантности документов, 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 [19].

В развитие работы планируется экспериментальная проверка эффективности предложенной методики.

Автор выражает благодарность И.П. Норенкову за постановку рассмотренной в работе задачи, а также за конструктивные обсуждения подходов к ее решению.

Работа выполнена при поддержке гранта РФФИ 10-07-00401.

Литература

1. И.П. Норенков. Интеллектуальные технологии на базе онтологий // Информационные технологии, 2010, №1, с.17-23.

2. The Dublin Core Metadata Initiative [Электронный ресурс]. (http://dublincore.org/).

3. А.П. Карпенко. Меры важности концептов в семантической сети онтологической базы знаний [Электронный ресурс] // Наука и образование: электронное научно- техническое издание, 2010, 7. (http://technomag.edu.ru/doc/151142.html).

4. G. Karypis, V. Kumar. Multilevel k-way Partitioning Scheme for Irregular Graphs // Journal of Parallel and Distributed Computing, 1998, vol. 8, no. 1, pp. 96-129.

5. Д.П. Бувайло, В.А. Толок. Быстрый высокопроизводительный алгоритм для разделения нерегулярных графов // Вісник Запорізького державного університету, 2002, № 2, с. 1 – 10.

6. T. N. Bui, S. Chaudhuri, F. T. Leighton, M. Sipser. Graph bisection algorithms with good average case behavior // Combinatorica, 1987, N7, pp. 171.191.

7. L. Miller Gary, Teng Shang-Hua, A. Vavasis Stephen. A unified geometric approach to graph separators: Proceedings of 31st Annual Symposium on Foundations of Computer Science, 1991, pp. 538 -547.

8. М.Р. Когаловский. Перспективные технологии информационных систем. – М.: ДМК Пресс; М.: Компания АйТи, 2003. – 288 с.

9. G.A. Miller and etc. Wordnet: a lexical database for the english language [Электронный ресурс]. // (http://wordnet.princeton.edu/).

10. E. Gabrilovich, S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07), Hyderabad, India, January 6-12, 2007: AAAI Press, 2007, pp. 1606–1611.

11. Ю.А. Целых. Теоретико-графовые методы анализа нечетких социальных сетей [Электронный ресурс]. (http://swsys.ru/print/article_print.php?id=742).

12. B. Hendrickson, R. Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. Sandia National Laboratories. - Technical Report SAND92-1460, 1992. –P. 192.

13. М. Гринева, Д. Лизоркин. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов [Электронный ресурс]. (http://citforum.ru/database/articles/kw_extraction/).

14. М.Ю. Богатырев, В.Е. Латов, И.А. Столбовская. Применение концептуальных графов в системах поддержки электронных библиотек: Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007. – Т. 2, С. 104-110.

15. Л.И. Бородкин. Математические методы и компьютер в задачах атрибуции текстов [Электронный ресурс]. (http://www.textology.ru/library/book.aspx?bookId=11&textId=13).

16. Dmitry Lizorkin and etc. Accuracy Estimate and Optimization Techniques for SimRank Computation: Proceedings of the 34th International Conference on Very Large Data Bases (VLDB'08). – 2008. – Vol. 1, Issue 1. – pp. 422-433;

17. О.И. Ларичев. Теория и методы принятия решений, а также Хроника событий в Волшебных странах. – М.: Университетская книга, Логос, 2006. -292 с.

18. Hong Zhang, Masumi Ishikawa. Evolutionary Canonical Particle Swarm Optimizer - A Proposal of Meta-Optimization in Model Selection. Berlin : Springer-Verlag, 2008.

19. R. Mihalcea. 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.