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

Меры важности концептов в семантической сети онтологической базы знаний

 # 07, июль 2010
автор: Карпенко А. П.

УДК 519.6

МГТУ им. Н.Э. Баумана,

Введение

Можно выделить три следующих класса систем поддержки принятия решений (СППР): системы, основанные на использовании типовых решений; системы, использующие типовые правила синтеза решений; системы, использующие поиск прецедентов. Корпоративная база знаний представляет собой, как правило, совокупность 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 СППР, которые используют такие базы знаний, поиск решения заключается в поиске в этих базах наиболее подходящих прецедентов и соответствующих им документов [1].

Эффективность поиска решений в базах знаний прецедентов в значительной мере зависит от используемых методов поиска.

Современные поисковые системы основаны, преимущественно, на применении полнотекстового поиска – поиска в каждом из документов всех

терминов, входящих в запрос. При этом учитывается частота встречаемость терминов в документе и их средняя языковая частотность [2].

Более эффективной альтернативой полнотекстовому поиску является поиск по метаданным – поиск по атрибутам документов, содержащимся в их метаданных. При этом 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].

Эффективный поиск решений в базах знаний прецедентов должен, очевидно, основываться не на регистрационных атрибутах документов, а на параметрах, характеризующих ситуацию принятия решения и само решение. Поэтому для СППР классический поиск по метаданным может играть лишь вспомогательную роль.

В работе рассматривается подход к поиску решений в базах знаний прецедентов, в котором метаданные формируются на основе онтологии соответствующей предметной области, заданной в виде семантической сети. При этом релевантность документов оценивается близостью в некоторой метрике концептов, входящих в метаданные документа, и концептов поискового запроса [1]. Можно предложить значительное количество таких метрик, при построении которых может оказаться целесообразным учитывать «важность» фигурирующих в них концептов.

В работе предлагается ряд мер важности концептов в семантической сети онтологической базы знаний. При разработке этих мер существенно используются некоторые результаты наших публикаций [4 - 8].

Модели семантических сетей

Пусть $\mathbf{c}(O) = \{c_i, i \in [1 : n_O]\}$ - множество концептов рассматриваемой онтологии O, а $\mathbf{r}(O) = \{r_j, j = 1,2,...\}$ - совокупность четких бинарных отношений между концептами множества $\mathbf{c}(O)$. Положим, что каждое из

бинарных отношений $\mathbf{r}(O)$ принадлежат к одному из типов отношений $R(O) = \{R_\alpha, \alpha \in [1 : m_O]\}$. Здесь n_O, m_O – общее число концептов онтологии O и число типов отношений, соответственно. Примерами типов семантических отношений являются таксономические, характеристические, каузальные, атрибутивные, квантифицирующие, временные, пространственные, арифметические, логические и многие иные типы отношений.

Семантическую сеть $S(O)$ онтологии O представим в виде взвешенного мультиграфа $G(O)$, вершины которого соответствуют концептам множества $\mathbf{C}$, а дуги – отношениям между ними. Заметим, что граф $G(O)$ не обязательно является связным.

Пусть $\mathbf{c}(T) \subset \mathbf{c}(O)$ - множество концептов рассматриваемого документа T, а $\mathbf{r}(T) \subset \mathbf{r}(O)$ - совокупность бинарных отношений между концептами множества $\mathbf{c}(T)$. Общее число концептов и типов отношений $R(T)$ документа T обозначим n_T, m_T соответсвенно; $n_T \le n_O$, $m_T \le m_O$.

Семантическую сеть $S(T) \subset S(O)$ документа T представим в виде взвешенного мультиграфа $G(T)$, аналогичного графу $G(O)$.

Веса вершин и дуг графов $G(O)$, $G(T)$ определены ниже.

Метрики графа семантических сетей

При построении мер важности концептов в семантических сетях $S(O)$, $S(T)$ используются рассмотренные ниже метрики соответствующих графов $G(O)$, $G(T)$.

Положим прежде, что между собой связаны все концепты семантической сети $S(O)$ и что отношения, связывающие эти концепты, являются отношениями частичного порядка типа R_α (например, родо-видовыми отношениями). Графа $G(O)$ в этом случае представляет собой ориентированный граф,

Тогда в качестве метрик графа $G(O)$ могут быть использованы его «высота», которая определяется на основе ярусно-параллельной формы (ЯПФ) этого графа [9].

Номер яруса ЯПФ графа $G(O)$, на котором находится концепт c_i, называется *высотой концепта* и обозначается $h_\alpha^O(c_i)$; количество ярусов в ЯПФ графа $G(O)$ называется *высотой графа* и обозначается $h_\alpha(G(O)) = h_\alpha^O$.

Положим теперь, что тип отношений R_α не принадлежит типу отношений частичного порядка. В этом случае в качестве метрики графа $G(O)$ может быть использован «диаметр графа» $a_\alpha(G(O)) = a_\alpha^O$, которым называется максимальное расстояние между его двумя вершинами. Расстоянием ρ между вершинами графа называется минимальное количество ребер графа, связывающих эти вершины [10].

Наконец, в качестве метрики графа $G(O)$ может быть использована его «реберная плотность», определяемая формулой

$$b_\alpha(G(O)) = b_\alpha^O = \frac{2\beta}{\gamma(\gamma-1)},$$

где β - количество дуг этого графа, а $\gamma = n_O$ - количество его вершин. Реберная плотность $b_\alpha^O \in [0,1]$ и характеризует близость графа $G(O)$ к полносвязному графу (клике): чем ближе величина b_α^O к единице, тем выше связность графа $G(O)$ и он ближе к полносвязному графу.

Аналогично, при построении мер важности концептов в семантической сети документа $S(T)$ используются следующие метрики графа $G(T)$: высота концепта $h_\alpha^T(c_i)$; высота графа $h_\alpha(G(T)) = h_\alpha^T$; диаметр графа $a_\alpha(G(T)) = a_\alpha^T$; реберная плотность $b_\alpha(G(T)) = b_\beta^T$.

Кластеризация семантических сетей

Если концепты c_i, c_j семантической сети $S(O)$ связаны между собой отношением типа $R_\alpha \in R(O)$, то будем говорить, что эти концепты связаны отношением типа R_α в узком смысле. Число всех концептов множества $\mathbf{c}(O)$, включая концепт c_i, связанных отношением типа R_α с этими концептом в узком смысле, обозначим $n_\alpha^O(c_i)$.

Пусть в узком смысле отношением типа R_α концепт c_i связан с концептом c_j, концепт c_j - с концептом c_k и так далее до концепта c_q. Здесь полагается, что все концепты c_i, $c_j,\ldots,c_q$ принадлежат множеству концептов $\mathbf{c}(O)$. Тогда будем говорить, что концепты c_i, c_q связаны отношением типа R_α в широком смысле. Число всех концептов семантической сети $S(O)$, включая концепт c_i, связанных отношением типа R_α с этим концептом в широком смысле, обозначим $N_\alpha^O(c_i)$.

Обозначим $d_\alpha^O(c_i)$ совокупность всех концептов семантической сети $S(O)$, включая сам концепт c_i, которые связаны отношением типа R_α с концептом c_i в узком смысле. Назовем эту совокупность R_α-*локальным кластером концепта* c_i в семантической сети $S(O)$. Число концептов в кластере $d_\alpha^O(c_i)$ равно, очевидно, $n_\alpha^O(c_i)$.

Отметим, что, поскольку концепт $c_i \in \mathbf{c}(O)$ может одновременно входить в несколько локальных кластеров, кластеры $d_\alpha^O(c_i)$, $d_\beta^O(c_i)$, $\alpha,\beta \in [1:m_O]$, $\alpha \neq \beta$ могут пересекаться, так что, вообще говоря,

$$d_\alpha^O(c_i) \bigcap d_\beta^O(c_i) \neq \varnothing .$$

Аналогично, обозначим $D_\alpha^O(c_i)$ совокупность всех концептов семантической сети $S(O)$, включая сам концепт c_i, которые связаны отношением типа R_α с концептом c_i в широком смысле, и назовем эту

совокупность R_α-*глобальным кластером концепта* c_i в семантической сети $S(O)$. Легко видеть, что число концептов в кластере $D_\alpha^O(c_i)$ равно $N_\alpha^O(c_i)$.

Отметим, что кластер $D_\alpha^O(c_i)$ является одновременно R_α-глобальным кластером всех концептов, принадлежащих этому кластеру.

Совокупность всех концептов кластера $D_\alpha^O(c_i)$, включая концепт c_i, которые расположены на расстоянии $\rho = 1,2,...$ от указанного концепта, обозначим $D_\alpha^O(\rho, c_i)$. Число таких концептов обозначим $N_\alpha^O(\rho, c_i)$, где $a(D_\alpha^O(c_i))$ - диаметр кластера $D_\alpha^O(c_i)$; $\rho = 1,2,...,a(D_\alpha^O(c_i))$. Очевидно, что $N_\alpha^O(1, c_i) = n_\alpha^O(c_i)$.

Взвешенные мультиграфы, соответствующие кластерам $d_\alpha^O(c_i)$, $D_\alpha^O(c_i)$ обозначим $g_\alpha^O(c_i)$, $G_\alpha^O(c_i)$ соответственно.

Аналогично определим связи концептов документа T в узком и широком смыслах, а также введем в рассмотрение R_α-локальный и R_α-глобальный кластеры $d_\alpha^T(c_i)$, $D_\alpha^T(c_i)$ документа T. Числа концептов в этих кластерах обозначим $n_\alpha^T(c_i)$, $N_\alpha^T(c_i)$ соответственно. Введем в рассмотрение также взвешенные мультиграфы $g_\alpha^T(c_i)$, $G_\alpha^T(c_i)$, аналогичные графам $g_\alpha^O(c_i)$, $G_\alpha^O(c_i)$. Кроме того, рассмотрим совокупности концептов $D_\alpha^T(\rho, c_i)$, аналогичные совокупностям $D_\alpha^O(\rho, c_i)$, где $\rho = 1,2,...,a(D_\alpha^T(c_i))$.

Веса вершин и дуг семантических сетей

Поставим в соответствие каждому из типов отношений $R_\alpha \in R(O)$ его вес v_α^O, $\alpha \in [1:m_O]$. Аналогично поставим в соответствие каждой из вершин $c_i \in \mathbf{c}(O)$ графа $G(O)$ вес w_i^O, $i \in [1:n_O]$, формализующий «важность» концепта c_i в семантической сети $S(O)$.

Определим прежде веса v_α^O, $\alpha \in [1:m_O]$. В простейшем случае в качестве веса v_α^O можно использовать общее число n_α^O концептов онтологии O, связанных между собой отношением типа R_α:

$$v_\alpha^O = n_\alpha^O.$$

Если отношения типа R_α представляют собой отношения частичного порядка, то в качестве веса v_α^O может быть использована максимальная из высот R_α-глобальных кластеров $D_\alpha^O(c_i)$:

$$v_\alpha^O = \mathbf{max}\ h\left(D_\alpha^O(c_i)\right),\ c_i \in \mathbf{c}(O).$$

Аналогично, вес v_α^O можно определить на основе суммарной и средней высот $h\left(D_\alpha^O(c_i)\right)$:

$$v_\alpha^O = \sum_i h\left(D_\alpha^0(c_i)\right),\ i \in [1:n_O];$$

$$v_\alpha^O = \frac{1}{n_O}\sum_i h\left(D_\alpha^0(c_i)\right),\ i \in [1:n_O].$$

Для произвольного типа отношений R_α в качестве веса v_α^O могут быть использованы максимальный, суммарный и средний диаметры соответствующих кластеров:

$$v_\alpha^O = \mathbf{max}\ a\left(D_\alpha^O(c_i)\right),\ c_i \in \mathbf{c}(O);$$

$$v_\alpha^O = \sum_i a\left(D_\alpha^0(c_i)\right),\ i \in [1:n_O];$$

$$v_\alpha^O = \frac{1}{n_O}\sum_i a\left(D_\alpha^0(c_i)\right),\ i \in [1:n_O].$$

Аналогично, веса v_α^O можно определить на основе максимальной, суммарной и средней реберных плотностей графов $D_\alpha^O(\rho,c_i)$, где ρ - некоторое фиксированное значение из диапазона $1,2,...,a(D_\alpha^T(c_i))$. Например,

$$v_\alpha^O(\rho) = \mathbf{max}\ b\left(D_\alpha^O(\rho,c_i)\right),\ c_i \in \mathbf{c}(O).$$

Наряду с рассмотренными весами v_α^O могут быть использованы их нормированные тем или иным образом аналоги, например,

$$v_\alpha^O = n_\alpha^O \Big/ n_O,\ v_\alpha^O = \mathbf{max}\ h\left(D_\alpha^O(c_i)\right) \Big/ h^O$$

и т.д. Большое количество выражений для весов v_α^O может быть получено на основе использования различных сверток рассмотренных весов.

Положим, что веса отношений $v_\alpha^O,\ \alpha \in [1:m_O]$ тем или иным образом определены. Тогда в простейшем случае в качестве веса w_i^O может быть использовано взвешенное число концептов, содержащихся во всех R_α-локальных кластерах $d_\alpha^O(c_i)$:

$$w_i^O = \sum_\alpha v_\alpha^O n_\alpha^O(c_i),\ \alpha \in [1:m_O].$$

Аналогично можно использовать взвешенное число концептов, содержащихся во всех R_α-глобальных кластерах $D_\alpha^O(c_i)$:

$$w_i^O = \sum_\alpha v_\alpha^O N_\alpha^O(c_i),\ \alpha \in [1:m_O]. \qquad (1)$$

Положим, что в формуле (1) влияние концептов на вес концепта c_i изменяется по мере увеличения расстояния этих концептов от концепта c_i, например, обратно пропорционально этому расстоянию. Тогда из формулы (1) следует формула

$$w_i^O = \sum_\alpha v_\alpha^O \sum_\rho \frac{1}{\rho} N_\alpha^O(\rho, c_i),\ \alpha \in [1:m_O],\ \rho \in \left[1 : a\left(D_\alpha^0(c_i)\right)\right].$$

Формулы для вычисления веса w_i^O могут быть построены на основе взвешенных максимального, суммарного и среднего из диаметров кластеров $D_\alpha^O(c_i)$:

$$w_i^O = \sum_\alpha \mathbf{max}\ v_\alpha^0 a\left(D_\alpha^O(c_i)\right),\ \alpha \in [1:m_O];$$

$$w_i^O = \sum_\alpha v_\alpha^0 a\left(D_\alpha^0(c_i)\right),\ \alpha \in [1:m_O];$$

$$w_i^O = \frac{1}{m_O}\sum_\alpha v_\alpha^0 a\left(D_\alpha^0(c_i)\right),\ \alpha \in [1:m_O].$$

Аналогично, веса w_i^O можно определить на основе максимальной, суммарной и средней реберных плотностей графа $D_\alpha^O(\rho, c_i)$, например

$$w_i^O(\rho) = \sum_\alpha \mathbf{max}\ v_\alpha^0 b\left(D_\alpha^O(\rho, c_i)\right),\ \alpha \in [1:m_O],\ \rho \in \left[1 : a\left(D_\alpha^0(c_i)\right)\right].$$

Наряду с рассмотренными весами v_α^O могут быть использованы их нормированные тем или иным образом аналоги, а также различные свертки этих весов.

Меры важности концептов в семантической сети документа

Положим, что вес w_i^O концепта c_i в онтологии O тем или иным образом определен; $i \in [1:n_O]$. Тогда в качестве меры μ_i^T важности концепта c_i в семантической сети $S(T)$ документа T могут быть использованы следующие меры.

1). Взвешенное число концептов, содержащихся во всех R_α-локальных кластерах $d_\alpha^T(c_i)$

$$\mu_i^T = \sum_\alpha w_\alpha^O n_\alpha^T(c_i),\ \alpha \in [1:m_O].$$

2). Взвешенное число концептов, содержащихся во всех R_α-глобальных кластерах $D_\alpha^T(c_i)$

$$\mu_i^T = \sum_\alpha w_\alpha^O N_\alpha^T(c_i),\ \alpha \in [1:m_O].$$

3). Нормированное взвешенное число концептов, содержащихся во всех R_α-глобальных кластерах $D_\alpha^T(c_i)$

$$\mu_i^T = \sum_\alpha w_\alpha^O \sum_\rho \frac{1}{\rho} N_\alpha^T(\rho, c_i),\ \alpha \in [1:m_O],\ \rho \in \left[1 : a\left(D_\alpha^T(c_i)\right)\right].$$

4). Взвешенные максимальный, суммарный и средний из диаметров кластеров $D_\alpha^T(c_i)$:

$$\mu_i^T = \sum_\alpha \mathbf{max}\ w_\alpha^0 a\left(D_\alpha^T(c_i)\right),\ \alpha \in [1:m_O];$$

$$\mu_i^T = \sum_\alpha w_\alpha^0 a\left(D_\alpha^T(c_i)\right),\ \alpha \in [1:m_O];$$

$$\mu_i^T = \frac{1}{m_O}\sum_\alpha w_\alpha^0 a\left(D_\alpha^T(c_i)\right),\ \alpha \in [1:m_O].$$

5). Взвешенные максимальная, суммарная и средняя из реберных плотностей кластера $D_\alpha^T(\rho, c_i)$, $\rho \in \left[1:a\left(D_\alpha^T(c_i)\right)\right]$:

$$\mu_i^T = \sum_\alpha \mathbf{max}\ w_\alpha^0 b\left(D_\alpha^T(\rho, c_i)\right),\ \alpha \in [1:m_O];$$

$$\mu_i^T = \sum_\alpha w_\alpha^0 b\left(D_\alpha^T(\rho, c_i)\right),\ \alpha \in [1:m_O];$$

$$\mu_i^T = \frac{1}{m_O}\sum_\alpha w_\alpha^0 b\left(D_\alpha^T(\rho, c_i)\right),\ \alpha \in [1:m_O].$$

Кроме того, в качестве мер важности можно, очевидно, использовать нормированные тем или иным образом аналоги рассмотренных мер важности, а также различные свертки тех же мер важности.

Отметим, что большие значения всех предложенных мер, соответствуют большим значениям важности соответствующих концептов.

Заключение

В работе под онтологий O понимается, так называемая, «легкая» онтология, определяемая парой вида $O = \langle \mathbf{c}, \mathbf{r} \rangle$, где **c** - множество концептов, а **r** - множество отношений между ними. В развитии работы планируется применить предложенную в ней методику оценки важности концептов к «тяжелой» онтологии, которая определяется тройкой $O = \langle \mathbf{c}, \mathbf{r}, \mathbf{f} \rangle$, где **f** - множество функций интерпретации, определенных на концептах и/или отношениях онтологии.

Под отношениями **r** в работе понимаются четкие отношения. Однако во многих случаях более адекватной является модель онтологии, в которой эти отношения понимаются как нечеткие. В этом случае возможен анализ важности концептов с учетом различий в «силе» связей между ними.

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. 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 // (http://dublincore.org/).

4. Карпенко А.П., Соколов Н.К. Оценка сложности семантической сети в обучающей системе // Наука и образование: электронное научно-техническое издание, 2008, 11, (http://technomag.edu.ru/doc/106658.html).

5. Карпенко А.П., Соколов Н.К. Расширенная семантическая сеть обучающей системы и оценка ее сложности // Наука и образование: электронное научно- техническое издание, 2008, 12, (http://technomag.edu.ru/doc/111716.html).

6. Карпенко А.П., Галямова Е.В., Соколов Н.К. Методика контроля понятийных знаний субъекта обучения в обучающей системе // Наука и образование: электронное научно- техническое издание, 2009, 2, (http://technomag.edu.ru/doc/115086.html).

7. Карпенко А.П., Соколов Н.К. Меры сложности семантической сети в обучающей системе // М.: Вестник МГТУ им. Н.Э. Баумана, серия «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 // М.: Вестник МАДИ (ГТУ), 2009, №2(17), с.82-86.

9. Федотов И.Е. Некоторые приемы параллельного программирования: Учебное пособие.- М.: Изд-во МГИРЭА (ГУ), 2008.- 188 с.

10. Евстигнеев В.А. Применение теории графов в программировании. –М.: Наука, 1985.-332 с.

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