

И. А. БЕССМЕРТНЫЙ

УПРАВЛЕНИЕ КОНТЕКСТОМ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Исследована проблема поиска данных в информационных системах, построенных на принципах Semantic Web. Обоснована необходимость использования данных контекста в поисковых запросах, предложена концептуальная модель интеллектуального агента, осуществляющего поиск с автоматическим управлением контекстом.

Ключевые слова: информационная система, контекст, онтология, информационный поиск.

Допущение открытого мира и поиск в базах знаний. Системы управления базами данных предоставляют возможность абстрагироваться от файловой системы, индексации и пр. при программировании приложений. При этом остается невозможным доступ на уровне языка SQL для конечного пользователя, поскольку для успешного поиска в базах данных (БД) при составлении SQL-запросов помимо владения языком SQL требуется знание структуры данных. Это означает, что запрос к БД всегда выполняется в заранее определенном и хорошо известном программисту контексте. Появление новых сущностей или отношений требует перепроектирования БД. Отсутствие в БД данных о факте автоматически означает его отрицание, т.е. в БД реализуется допущение замкнутого мира.

В отличие от БД в базах знаний (БЗ) отсутствуют заранее определенные модели данных, а разные фрагменты БЗ могут храниться на распределенных сетевых ресурсах, объединяемых в виде семантической сети во время выполнения поискового запроса. Кроме того, в БЗ обычно применяется допущение открытого мира, при котором отсутствие данных о факте не означает отрицания факта. Указанные свойства БЗ приводят к тому, что вычисление любых, даже очевидных, фактов может потребовать глобального поиска, сложность которого будет несоизмерима ценности полученного результата.

Из первых двух свойств вытекает третье, связанное с формированием запросов к БЗ. Аналогом инфологической модели БД для баз знаний является онтология предметной области. Поисковый запрос к БЗ должен базироваться на онтологии, а интерпретация онтологии представляет собой сложную задачу. Между тем допущение открытого мира означает необходимость неограниченного расширения пространства поиска, а значит и автоматической интерпретации онтологий. В рамках настоящего исследования предполагается, что поиск выполняется в документах, подготовленных на основе одной и той же онтологии.

Роль контекста в задаче поиска. Извлечение знаний из конкретного документа, созданного на основе известной онтологии, практически не отличается от обработки SQL-запросов к БД. Если документ не задан в запросе, возникает задача его поиска. Поиск с помощью серверов, ориентированных на использование человеком (Google, Yahoo, Yandex и др.), бесполезен,

поскольку теги HTML и любой другой разметки не индексируются, в то время как знания формализуются с применением разметки. В последнее время появился ряд сервисов, таких как SWOOGLE <<http://swoogle.umbc.edu/>>, SWSE <<http://swse.org/>> и др., позволяющих осуществлять поиск формализованных данных. Необходимость извлечения документов интеллектуальным агентом, в том числе с использованием таких поисковых сервисов, делает актуальной задачу автоматической оценки релевантности найденных документов запросу.

При написании документа всегда предполагается, что читатель знаком с его контекстом. Если название документа не полностью отражает содержание, то оно уточняется во вводной части. В лингвистике контекст есть фрагмент текста минус определяемая единица [1]. Таким образом, контекст $\xi(m)$ для сообщения m устанавливается как

$$\xi(m) = \langle C, L, P, I, V \rangle - m,$$

где C — множество классов объектов, L — множество отношений или предикатов (связей) между объектами, P — множество свойств объектов, I — множество экземпляров объектов, V — множество значений. Следовательно, для каждого последующего m_{i+1} сообщения предыдущее m_i включается в состав контекста:

$$\xi(m_{i+1}) = \xi(m_{i-1}) + m_i.$$

Данное определение является антропоморфным, ориентировано на речевое взаимодействие и предполагает, что субъект и объект такого взаимодействия обладают интеллектом и могут идентифицировать (установить) контекст для каждого коммуникационного акта. В случае информационного поиска установление контекста может быть простым только для локальных БЗ, как, например, это делается с помощью микротеорий в продукте ResearchCyc компании Cycorp <www.cyc.com>. Автор запроса к БЗ должен знать, каким образом факты в базе знаний группируются в микротеории, и эксплицитно указывать идентификатор микротеории в запросе. Необходимость владения микротеориями, идентификаторами элементов БЗ, а также специальным языком запросов делает взаимодействие с такой БЗ подобным работе с БД, которая доступна только для программиста, а не конечного пользователя.

В настоящей работе под контекстом будем понимать множество понятий предметной области, которое должно быть общим для всех участников информационного обмена и позволять обмениваться короткими сообщениями. Другое назначение контекста — ограничение предметной области, позволяющее сократить размерность задачи поиска и избежать противоречивости фактов. Обычно контекст задается в начале коммуникационного акта и при необходимости может уточняться. Таким образом, для правильной интерпретации сообщений все участники обмена информацией должны владеть контекстом. Визуализация знаний также требует использования контекста, как это сделано, в частности, в разработанной автором программе Semantic [2, 3], предназначенной для создания БЗ, извлечения и визуализации знаний.

Концептуальная модель агента для поиска в базах знаний. В соответствии с концепцией Semantic Web [4] извлечение знаний осуществляется интеллектуальными агентами, которые самостоятельно отыскивают требуемую информацию, формулируя при необходимости запросы к другим агентам. В отличие от SQL-запросов к БД (SELECT ... FROM ... WHERE ...) запросы к БЗ, в частности на языке SPARQL, не содержат конструкции FROM, поскольку запрос всегда выполняется в пределах целого документа. Поиск документа должен выполняться по контексту запроса. Это означает, что извлечение фактов из Semantic Web представляет собой двухэтапный процесс (рис. 1).

Поскольку точное совпадение контекста документа и контекста запроса является идеальным случаем, данный процесс не всегда завершается успешно и может итеративно повторяться.

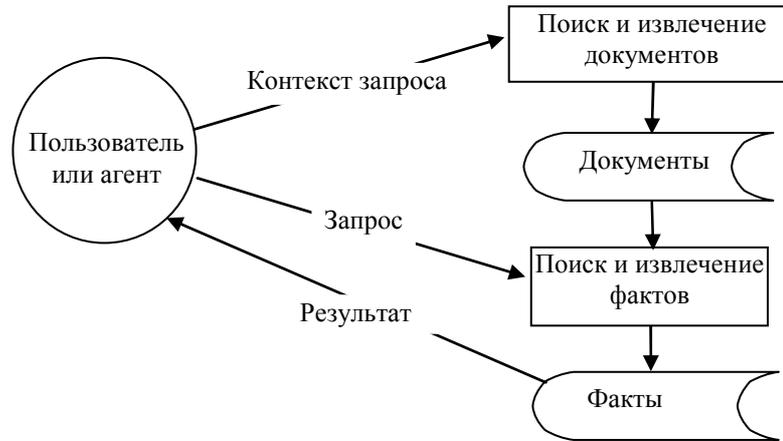


Рис. 1

На рис. 2 показаны все возможные сочетания контекста запроса Q и контекста предметной области (домена) D :

- a) запрос породил набор фактов, ни один из которых не является релевантным (отсутствие решений);
- b) найдены часть релевантных фактов и некоторое количество нерелевантных (решение есть, но оно неполное и противоречивое);
- c) найдены все релевантные факты и часть нерелевантных (решение полное, но противоречивое);
- d) найдена часть релевантных фактов и при этом нет нерелевантных (решение непротиворечивое, но неполное);
- e) найдены все релевантные факты и ни одного нерелевантного (целевое состояние: решение полное и непротиворечивое).

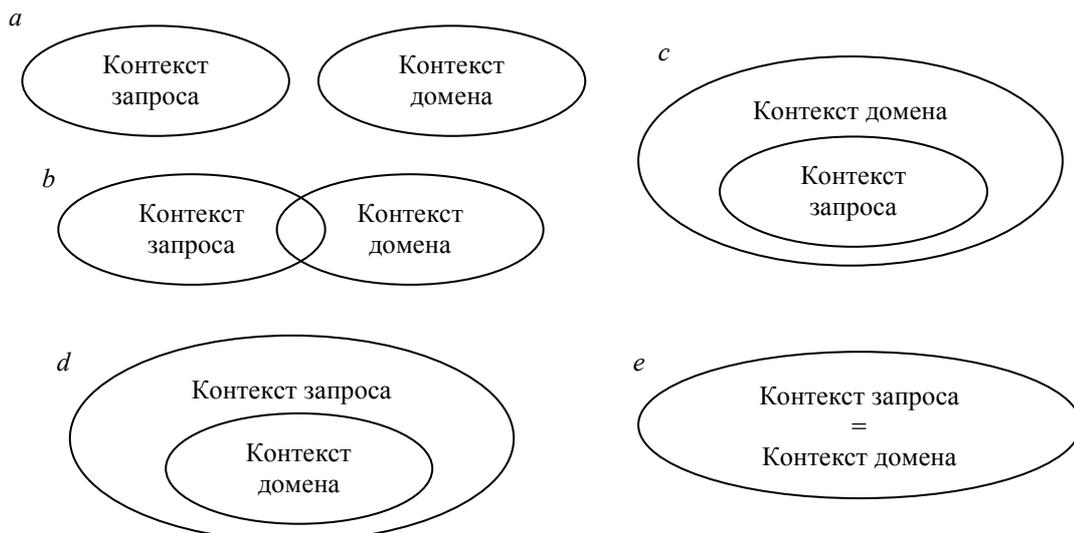


Рис. 2

В таблице приведены условия появления данных сочетаний и признаки, с помощью которых можно идентифицировать каждое из состояний.

Идентификатор	Решение			Условие
	наличие	полнота	противоречивость	
<i>a</i>	0	0	0	$Q \cap D = \emptyset$
<i>b</i>	1	0	1	$Q \cap D \neq \emptyset \ \& \ Q \cap D \neq Q \cup D$
<i>c</i>	1	1	1	$Q \subset D$
<i>d</i>	1	0	0	$Q \supset D$
<i>e</i>	1	1	0	$Q = D$

Если запрос не привел к состоянию *e*, следует изменить запрос. Ограничим модификации запросов двумя операциями: *x* — обобщение запроса (расширение контекста) и *y* — уточнение запроса (сужение контекста). Граф на рис. 3 отображает конечный автомат, в котором разрешенными являются только переходы, приближающие к целевому состоянию *e* (не отдаляющие от *e*).

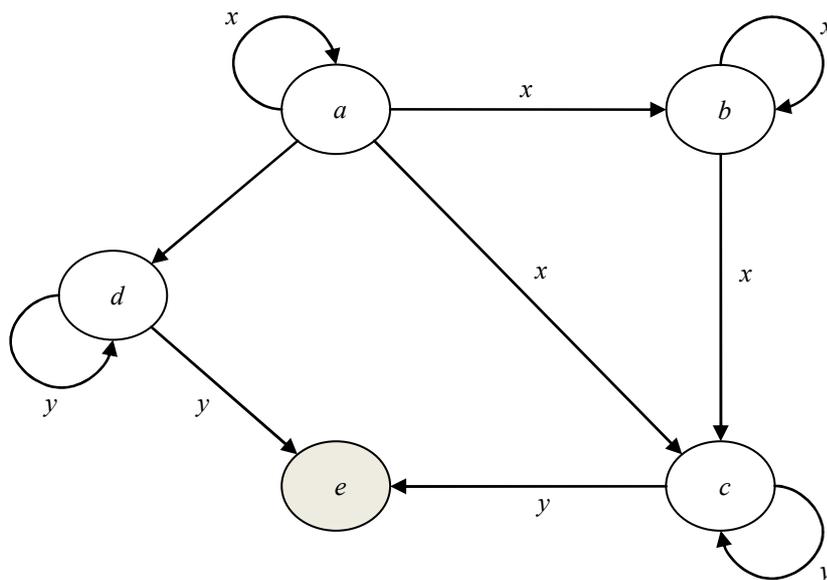


Рис. 3

Расширение пространства поиска может привести к увеличению времени извлечения фактов за счет применения правил, и это обстоятельство требует исследования.

Исследование производительности поиска в расширяющемся контексте. Пусть *N* — число фактов в базе знаний, τ — среднее время доступа к одному факту, *A_i* — число атрибутов или классов, релевантных *i*-му условию запроса, $i=1, n$. При обработке запроса первое условие применяется в среднем к $A_1/2$ атрибутам (классам), второе — если первое выполнено, а вероятность данного события $p_1=1/A_1$ и т.д. Заметим, что речь идет не о ветвлении поиска на дереве решений, а о фильтрации фактов, поэтому здесь нет экспоненциального роста сложности поиска.

Таким образом, среднее время *T* обработки запроса на извлечение фактов для точно определенного контекста составит

$$\begin{aligned}
 T &= \tau N \left(1 + \frac{A_1}{2} + p_1 \frac{A_2}{2} + p_1 p_2 \frac{A_3}{2} + \dots + p_1 p_2 \dots p_{n-1} \frac{A_n}{2} \right) = \\
 &= \tau N \left(1 + \frac{A_1}{2} + \frac{A_2}{2A_1} + \frac{A_3}{2A_1 A_2} \dots + \frac{A_n}{2A_1 A_2 \dots A_{n-1}} \right) = \tau N \left(1 + \frac{1}{2} \sum_{i=1}^n \frac{A_i}{\prod_{k=1}^{i-1} A_k} \right).
 \end{aligned}$$

Для $A=A_1=A_2=\dots=A_n$ при $n \geq 2$

$$T = \tau N \left(1 + \frac{A}{2} + \frac{1}{2} + \frac{1}{2A} + \dots + \frac{1}{2A^{n-1}} \right) \approx \tau N \frac{(A+3)}{2},$$

если пренебречь малыми значениями членов ряда, заключенных в скобки.

В случае поиска в расширенном контексте для каждого условия запроса из базы знаний должны извлекаться факты с близкими значениями атрибутов. Пусть δ_i — число допустимых значений i -го атрибута. При этом для каждого атрибута в условии запроса число извлекаемых фактов удваивается, поскольку требуются дополнительные факты, характеризующие близость значений атрибута. Тогда среднее время T_x обработки запроса на извлечение фактов для расширенного контекста будет

$$T_x = \tau N \left(1 + A_1 + \frac{\delta_1 A_2}{A_1} + \frac{\delta_1 \delta_2 A_3}{A_1 A_2} + \dots + \frac{\delta_1 \delta_2 \dots \delta_{n-1} A_n}{A_1 A_2 \dots A_{n-1}} \right) = \tau N \left(1 + A_1 + \sum_{i=1}^n A_i \prod_{j=1}^{i-1} \frac{\delta_j}{A_j} \right).$$

Для случая, когда $A=A_1=A_2=\dots=A_n$ и $\delta=\delta_1=\delta_2=\dots=\delta_n$, при $n \geq 2$,

$$T_x = \tau N \left(1 + A + \delta + \frac{\delta^2}{A} + \dots + \frac{\delta^{n-1}}{A^{n-2}} \right) \approx \tau N (1 + A + \delta),$$

если также пренебречь малыми значениями членов ряда.

Если контекст расширяется до целого класса, то для i -го условия запроса атрибут может принадлежать одному из C_i классов. Первое условие запроса порождает в среднем A_1+C_1 обращений к базе знаний. Вероятность успешного выполнения первого условия $p_1=1/C_1$. Тогда среднее время T_c обработки запроса на извлечение фактов в пределах класса для каждого из условий запроса составит

$$T_c = \tau N \left(1 + A_1 + C_1 + \frac{A_2 + C_2}{C_1} + \frac{A_3 + C_3}{C_1 C_2} + \dots + \frac{A_n + C_n}{C_1 C_2 \dots C_{n-1}} \right) = \tau N \left(1 + \sum_{i=1}^n \frac{A_i + C_i}{\prod_{j=1}^{i-1} C_j} \right).$$

Для случая, когда $A=A_1=A_2=\dots=A_n$ и $C=C_1=C_2=\dots=C_n$,

$$T_c = \tau N \left(1 + A + C + \frac{A}{C} + 1 + \frac{A}{C^2} + \frac{1}{C} + \dots + \frac{A}{C^n} + \frac{1}{C^{n-1}} \right).$$

Экспериментальное исследование скорости извлечения фактов производилось в среде SWI-Prolog на синтетической базе фактов, фрагмент которой представлен ниже.

t(1, isa, mathematician).	t(1, lives_in, denmark).
t(2, isa, player).	t(2, lives_in, korea).
t(3, isa, restorer).	t(3, lives_in, mexico).
t(4, isa, graver).	t(4, lives_in, spain).
t(5, isa, informatics).	t(5, lives_in, usa).
t(6, isa, conductor).	t(6, lives_in, switzerland).
...	
t(painter, isa, artist).	t(painter, id, 1).
t(graphic_artist, isa, artist).	t(graphic_artist, id, 2).
...	
t(matematician, isa, scientist).	t(matematician, id, 13).
...	
t(conductor, isa, musician).	t(conductor, id, 23).
t(singer, isa, musician).	t(singer, id, 24).
...	
t(korea, isa, asia).	t(korea, id, 1).
t(china, isa, asia).	t(china, id, 2).
...	
t(england, isa, europe).	t(england, id, 12).
t(usa, isa, america).	t(usa, id, 13).

Базу составляют сгенерированные случайным образом факты о субъектах, идентифицируемых числами и имеющих атрибуты „профессия“ и „страна“. Экземпляры группируются в классы (континент, ученый, художник, инженер, ...). Кроме того, атрибуты имеют численные идентификаторы, присвоенные таким образом, чтобы близкие значения идентификаторов соответствовали близким профессиям или соседним странам. В точном запросе выбирались все представители конкретной профессии, проживающие в конкретной стране, в расширенном запросе — профессии и страны, имеющие смежные идентификаторы, а в поиске по классам — субъекты, представляющие класс профессий и континент.

На рис. 4 представлены теоретические результаты (цифры без штриха) и данные, полученные с помощью экспериментов (цифры со штрихом) для поиска в точно заданном контексте 1, а также поиска в ближайшем окружении заданного контекста 2 и в классе, объединяющем все контексты уровнем выше 3. Приведенные результаты демонстрируют, во-первых, хорошее совпадение теоретических результатов с экспериментами, во-вторых, линейный рост сложности поиска при увеличении числа фактов в базе данных, в-третьих, заметно большее время поиска в ближайшем окружении заданного контекста по сравнению с поиском в пределах целого класса. Последнее обстоятельство объясняется тем, что при поиске в целом классе делается меньше проверок условий.

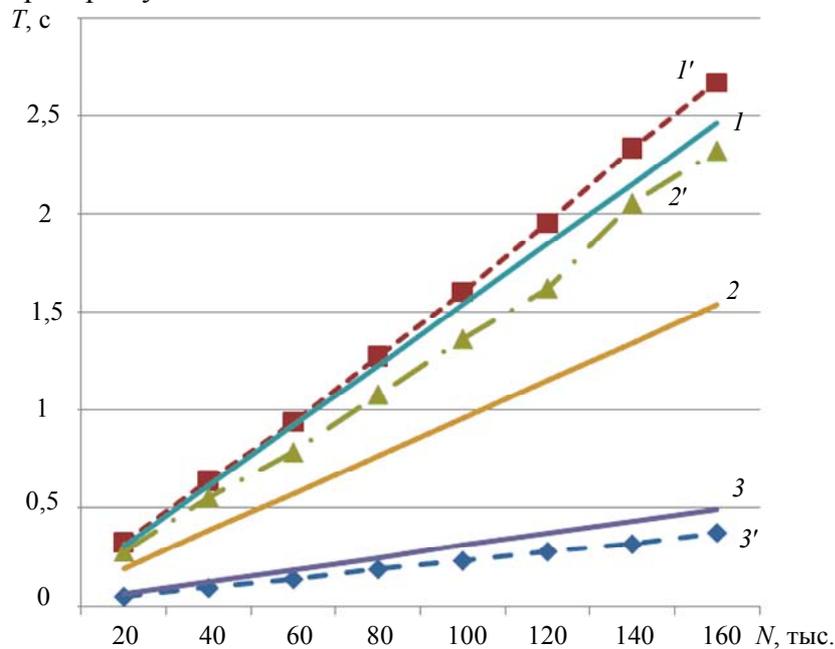


Рис. 4

Заключение. Представленные результаты исследования демонстрируют линейный рост сложности поиска в расширяющемся контексте, что позволяет использовать предложенную концептуальную модель интеллектуального агента для извлечения знаний из Semantic Web. Отдельного исследования заслуживает проблема идентификации контекста, решение которой может позволить автоматически оценивать степень доверия к результатам поиска.

Работа выполнена при финансовой поддержке ФЦП „Научные и научно-педагогические кадры инновационной России на 2009—2013 годы“ (соглашение № 14.В37.21.0406).

СПИСОК ЛИТЕРАТУРЫ

1. Торсуева И. Г. Контекст // Лингвистический энциклопедический словарь. М.: СЭ, 1990. С. 238—239.
2. Бессмертный И. А. Методы поиска информации с использованием интеллектуального агента // Изв. вузов. Приборостроение. 2009. Т. 52, № 12. С. 26—31.

3. *Bessmertny I. A.* Knowledge Visualization Based on Semantic Networks // Programming and Computer Software. 2010. Vol. 36, N 4. P. 197—204.
4. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web // Scientific American Magazine. 2001. May.

Игорь Александрович Бессмертный —

Сведения об авторе

канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники;
E-mail: igor_bessmertny@hotmail.com

Рекомендована кафедрой
вычислительной техники

Поступила в редакцию
08.02.12 г.