

В.А. Жигалов, Л.Г. Романов

РосНИИ Искусственного Интеллекта

zhigalov@aha.ru

Аннотация

На основе семантически-ориентированного подхода к анализу ЕЯ предложен метод оценки соответствия смысла ЕЯ-запроса определенной предметной области. Описаны основные принципы вычисления функции семантической корреляции (ФСК). Предложен способ вычисления ФСК на основе введения двух числовых параметров компонентов анализа – семантической плотности и семантического объема.

1. ЦЕЛИ

Темой настоящей статьи является задача определения соответствия короткого естественно-языкового текста определенной предметной области (ПО). Приложения этой задачи могут быть весьма разнообразными. В частности, одним из приложений может быть распределенный семантически-ориентированный поиск в Интернет, когда одна поисковая машина может перенаправлять поступивший запрос произвольному количеству других. Для того, чтобы оценить целесообразность перенаправления конкретного запроса конкретной поисковой машине, надо оценить вероятность получения релевантного ответа, один из способов этого - соотнести предметную область поисковой машины (поискового домена) и предметную область запроса.

2. ФУНКЦИЯ СЕМАНТИЧЕСКОЙ КОРРЕЛЯЦИИ - ОСНОВНЫЕ СВОЙСТВА

Для оценки и принятия решения удобно использовать числовые данные, поэтому эту оценку можно реализовать как функцию. Примем, что значение этой функции равно 1, если запрос максимально соответствует ПО домена, и 0, если семантика запроса независима (ортогональна) по отношению к ПО. Функция непрерывна и может принимать любые значения в интервале $[0, 1]$.

Эту функцию назовем *функцией семантической корреляции* (ФСК) [Жигалов 1999]. Функция семантической корреляции оценивает вероятность получения релевантного ответа на конкретный запрос доменом (т.е. отвечающей системой, например, сервером поиска или ЕЯ-интерфейса к базе данных) с определенной предметной областью.

Рассмотрим основные свойства ФСК, вытекающие из ее определения, и постулируем некоторые дополнительные. Функция зависит от запроса и предметной области. Очевидно, запрос должен быть представлен как последовательность символов. Относительно предметной области такой однозначности нет. Будем предполагать, что модель ПО представлена по крайней мере как механизм вычисления ФСК. Это означает, во первых, что как таковое явное структурное описание ПО в виде ее модели (представленной диаграммой сущность-связь, семантической сетью и т.д.) может отсутствовать, а во-вторых, ФСК в общем случае является алгоритмом либо процессом, и совершенно не обязательно должна иметь представление в виде явной зависимости (формулы).

Если домен содержит сразу несколько определенных четко различимых предметных областей и его ПО является их объединением, то ФСК для него вычисляется в зависимости от ФСК этих ПО:

$$(1 - r) = P(1 - r_i), \quad (1)$$

где r - результирующее значение функции семантической корреляции,
 r_i - значение ФСК i -ой предметной области.

Эта формула становится понятной, если учесть, что ФСК моделирует вероятность.

Если по какой-либо причине необходимо дополнительно ограничивать запросы, поступающие на данный домен, на ФСК накладывается фильтр, который представляет собой множитель, зависящий от запроса. Очевидно, этот множитель сам является ФСК, иными словами, результирующее значение может быть произведением нескольких различных функций:

$$r = Pr_i$$

(2)

Так вычисляется функция семантической корреляции для поисковых доменов, предметная область которых является пересечением нескольких ПО.

Проиллюстрируем это на примере поискового домена, который имеет предметную область книжного магазина. Если нам необходимо принимать запросы только про книги по экономической тематике, можно умножать "книжную" ФСК на ФСК, моделирующую экономическую предметную область. Этот второй множитель будет фильтром, отсекающим все "неэкономические" запросы.

3. СЕМАНТИЧЕСКИ-ОРИЕНТИРОВАННЫЙ ПОДХОД

Одним из эффективных способов анализа ЕЯ применительно к определенной предметной области является семантически-ориентированный подход [Нариньяни 1979].

Рассмотрим возможность применения семантически-ориентированного подхода к анализу ЕЯ-запросов для вычисления значения ФСК. Одной из особенностей этого подхода является построение структуры (сети или дерева) формального представления запроса методом "снизу вверх", при этом элементы (компоненты) дерева обычно получаются на основе низлежащих, более элементарных компонентов. При этом компоненты могут удаляться по мере построения на их основе других, а могут оставаться и связываться с другими компонентами, образуя таким образом структуру дерева формального представления запроса.

Самыми элементарными компонентами выступают лексемы запроса. Порождение компонентов на основе исходных может происходить по продукционной программе, а может быть организовано как параллельный недетерминированный процесс, реализующий концепции управления по данным и по событиям [Загорулько 1999], [Кононенко 1994]. Этот подход зарекомендовал себя как надежный способ понимания ЕЯ в ограниченных предметных областях. Заметим, что наша задача - оценка соответствия ЕЯ-запроса определенной ПО - весьма схожа с задачей семантически-ориентированного подхода к пониманию ЕЯ.

4. ПРИМЕНЕНИЕ СЕМАНТИЧЕСКИ-ОРИЕНТИРОВАННОГО ПОДХОДА К ВЫЧИСЛЕНИЮ ФСК

Если спроектировать этот подход на нашу задачу, представляется возможным организовать процесс вычисления ФСК следующим образом. Каждый компонент имеет определенный набор параметров, в том числе и числового типа. При создании компонента значения его параметров вычисляются в зависимости от значений параметров исходных компонентов, а также от значений параметров контекстных компонентов. В конце процесса создается самый верхний в дереве запроса компонент, и вычисляется само значение ФСК как функция параметров - компонентов дерева.

Для процесса вычисления ФСК требуется семантически-ориентированный словарь, с присвоением словам семантических типов, а также параметров (атрибутов), правила построения компонентов и вычисления их атрибутов на основе некоторых исходных, и модель предметной области, как вспомогательная информация для процесса анализа.

ОБОСНОВАНИЕ ВЫБРАННОГО ПОДХОДА

В описываемом подходе выполняется одновременное построение дерева и вычисление скаляра ФСК. Дерево запроса отражает семантику запроса в данной ПО, и эту семантику удобно описывать в сетевом представлении (в частном случае деревом). Весьма естественно представлять ФСК как функцию, зависящую от наличия либо отсутствия отдельных компонентов запроса, представленных узлами сети или дерева, а также от значений их параметров, поскольку все они эффективно описывают семантику запроса, а наша функция оценивает общее соответствие запроса данной ПО.

5. ВАРИАНТ РЕАЛИЗАЦИИ ВЫЧИСЛЕНИЯ ФСК - ДВУХАТРИБУТНАЯ МОДЕЛЬ

5.1 Семантическая плотность и семантический объем

Рассмотрим один из возможных вариантов реализации описываемого подхода. Каждому семантическому компоненту присвоим два атрибута, которые условно назовем *семантическая плотность* и *семантический объем* (смысл таких названий будет объяснен далее). Значение ФСК запроса в самом простом случае вычисляется как сумма «семантических масс» компонентов, деленная на сумму их объемов (то есть семантическая плотность запроса):

$$r = \sum r_i V_i / \sum V_i \quad (3)$$

где r_i – семантическая плотность компонента,
 V_i – семантический объем компонента.

Для того, чтобы ФСК была нормирована, максимальная плотность должна быть равной 1, минимальная - 0. Наглядной метафорой для такой модели может служить, например, соотношение каждому компоненту некоторого сосуда, с определенным объемом и некоторой степенью заполнения жидкостью единичной плотности.

Такой вариант вычисления ФСК позволяет учитывать не только важность данного компонента, но и важность его наличия и отсутствия по отношению к наличию и отсутствию других компонентов. Например, в число семантических типов слов, согласно семантически-ориентированному подходу, входят значимые и незначимые типы, а также неизвестные слова, которые могут составить отдельный семантический тип. Наличие незначимых, но известных слов должно незначительно влиять на ФСК, если в запросе есть значимые слова, и приводить к малому значению ФСК, если значимых слов нет. Неизвестное же слово значительно уменьшает семантическую релевантность запроса даже при наличии значимых слов. Значимые компоненты, напротив, должны увеличивать ФСК, причем увеличение это должно быть в зависимости от типа компонента. Для того, чтобы реализовать такое поведение ФСК, присвоим незначимым компонентам нулевую плотность и небольшой объем, компоненту неизвестной лексики - нулевую плотность и большой объем, и семантически значимым компонентам - большую плотность и большой объем. Тогда наличие значимой лексики будет увеличивать значение ФСК, и это увеличение будет тем сильнее, чем меньше неизвестной лексики в запросе.

Приведенная формула, безусловно, не позволяет достаточно точно моделировать поведение ФСК – она лишь раскрывает основной принцип. Если процесс построения сетевого представления запроса происходит с созданием дополнительных семантических компонентов (например, предикатов, их логических связей и т.д.), то эти компоненты также получают значения плотности и объема, в зависимости от соответствующих значений атрибутов исходных компонентов, по приведенной выше элементарной формуле, либо по более сложному алгоритму. В конце концов самый верхний элемент дерева получает значения ρ и V , и ρ представляет значение ФСК.

Очевидно, неизвестная лексика не образует производных компонентов, и в итоге на значение ФСК окажут влияние плотность и объем значимых и производных от них компонентов, и компонентов неизвестной лексики. В случае, если требуется поставить некоторые слова в разряд "нежелательных", то есть приводящих к гарантированно низкому значению функции семантической корреляции, то для таких слов можно установить нулевую плотность и очень большой объем. В этом случае большой объем будет служить "балластом" для всего запроса, и даже наличие значимой лексики не исправит положения.

5.2 Пример вычисления ФСК

На приведенном ниже рисунке показано дерево семантического анализа запроса «издательства, выпускающие компьютерную литературу». Лексемы (первоначальные, самые

элементарные компоненты) имеют типы - Атрибут (Attr), Значение (Val), Незначимое (Unsign), МаркерДомена (Domain). Моделируя процесс понимания запроса ЕЯ-интерфейсом к базе данных, строим дерево запроса, в котором присутствуют производные компоненты Предикат (Pred), Получить (Get), Дано (Cond), Запрос (Query). Из исходных компонентов наибольший вес имеет маркер предметной области "литература", наименьший - незначимое "выпускает".

Предположим, что плотность и объем компонентов Get и Cond зависят от параметров атрибутов (Attr) и предикатов (Pred), входящих в эти компоненты, следующим образом:

$$r = 1 - P(1 - r_i) \quad (4)$$

$$V = SV_i \quad (5)$$

То есть плотность этих компонентов возрастает с количеством подчиненных компонентов, что свидетельствует о более сложном осмысленном формальном представлении запроса, и, следовательно, большей вероятности принадлежности запроса нашей ПО. В рассматриваемом примере плотность и объем компонента Get будет равным плотности и объему атрибута (поскольку он один), то же - по отношению к компоненту Cond. Плотность компонента Query вычисляется по формуле (3). В результате значение ФСК равно 0.896.

Если же рассмотреть запрос "Цена и издательство книг по С++ или ООП" (рисунок ниже), то значение ФСК будет равным 0.95, за счет того, что у компонентов Cond и Get по два подчиненных компонента (см. формулы 4, 5). В этом запросе также есть маркер ПО - "книг", и незначимое "по". Рядом с компонентами показаны их плотность и объем. В запросе присутствуют также логические связки *и* и *или*, которые по своим характеристикам близки к незначимым, но выполняют служебную функцию при формировании составных компонентов Get и Cond из атрибутов и предикатов.

Заметим, что нет нужды присваивать значения параметров каждому слову в отдельности, достаточно расписать параметры для элементарных семантических типов, параметры же производных компонентов вычисляются в результате анализа ЕЯ-запроса.

5.3 Смысл плотности и объема компонентов - активная и пассивная значимости

Плотность и объем характеризуют два вида значимости компонента. Назовем первый вид - активной значимостью, а второй - пассивной. Активная значимость пропорциональна семантическому наполнению компонента, его "осмысленности" для данной ПО - она близка к 1 для семантически значимых, и к 0 для незначимых и неизвестных слов, а пассивная характеризует меру влияния активной значимости компонента на значимость всего запроса, то есть на значение ФСК, иными словами, важность компонента. Действительно, в самом простом случае, когда запрос состоит из одного слова, значение ФСК будет равно значению атрибута «плотность» этого слова, то есть пропорционально осмысленности этого слова для данной ПО. Пассивная же значимость является весовым коэффициентом перед значением активной значимости в формуле (3).

ВЫВОДЫ

1. Предлагаемый подход опирается на следующую гипотезу - *для достоверной оценки степени соответствия ЕЯ-запроса определенной предметной области необходимо выполнить задачу понимания запроса применительно к данной ПО.*
2. Поскольку вычисление ФСК сопровождается, по сути, семантически-ориентированным анализом запроса, можно довольно гибко манипулировать значением функции, и в результате довольно точно моделировать вероятность получения релевантного ответа на данный запрос, оценивая соответствие его семантики предметной области отвечающего домена.
3. Вычисление ФСК с помощью метафоры плотности и объема семантических компонентов запроса позволяет просто и наглядно моделировать поведение функции.

Литература

Жигалов В.А. Поиск в Интернет: от словарного индексирования к распределенной сети знаний. Труды международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Таруса 1999. стр. 47 - 56.

Загорулько Ю., Кононенко И., Попов И. Экспериментальная система понимания метеорологических телеграмм. // Труды международного семинара Диалог-99 по компьютерной лингвистике и ее приложениям., том 2, с. 57-66.

Кононенко И. Экспериментальная система понимания коротких текстов. // Труды IV национальной конференции ИИ-94 - Рыбинск, 1994 - с. 395-398

Нариньяни А.С., Лингвистические процессоры ЗАПСИБ (1-я и 2-я части). Препринт ВЦ СО АН СССР, N 199, 1979