Поисковая консультирующая система образовательного учреждения на основе семантической сети

Бородин В.А., Щербатов И.А.

***Аннотация*— В статье рассмотрена текущая тенденцию анализа поисковых запросов с точки зрения семантики выражения. На примере системы консультирования посетителей учебного заведения описывается модель работы подобной системы, основанная на семантическом подобии. В статье описывается алгоритм обработки одного слова и словосочетаний, а также описывается реализация прототипа системы. Пример, показанный в работе, показывает эффективность предложенных решений.**

***Ключевые слова*— семантическое подобие, семантическая сеть, системы консультирования.**

I. Введение

В сфере информационных технологий большое распространение получили системы, обрабатывающие запросы пользователей с точки зрения семантики введённого выражения. Такой подход позволяет улучшить результаты обработки запроса, например, исключив полисемию (одно слово имеет несколько значений) или синонимию (несколько слов с одним значением) [1]. Семантические технологии активно используются в алгоритмах информационного поиска. Например, семантику слов учитывает метод латентно-семантического индексирования. Этот метод, используя сингулярное разложение матриц, позволяет вычислить взаимосвязь между документами и терминами, сопоставляя их тематику [2]. Другим подходом анализа семантики выражения является построение семантических сетей. Семантические сети также получили большое распространение в области информационных технологий и информационного поиска. Многие поисковые движки, такие как Google, Kosmix, Powerset, DuckDuckGo используют семантические технологии. Например, среди результатов поискового запроса Google присутствует информация из семантической сети Google Knowledge Graph. Google Knowledge Graph основан на базе данных Freebase. Это большая база знаний, содержащая собранные сообществом пользователей данные, соединённые в виде семантической сети.

Другая активно используемая технология – это электронные словари, в которых слова соединяются семантическими связями. Ярким примером такого словаря является WordNet — семантическая сеть для английского языка. Словарь состоит из 4 сетей для основных частей речи: существительных, глаголов, прилагательных и наречий. Базовой словарной единицей в WordNet является синонимический ряд («синсет»), объединяющий слова со схожим значением и по сути своей являющийся узлом сети. Синсеты в WordNet могут быть связаны между собой различными семантическими отношениями: гиперонимами, гипонимами, синонимами, антонимами, холонимами, меронимами и др. [3]

Таким образом, объединение системы распознавания речи и последующего интеллектуального поиска позволяют создавать простые и удобные для пользователя информационные системы, которые могут быть применены в различных областях: например, в качестве мобильного ассистента или справочной системы учебного заведения. На данный момент в учебных заведениях накапливается большое количество информации. Расписание занятий, информация о различных мероприятиях, конференциях - все эти знания предназначены для широкого круга людей, однако они рассредоточены в различных источниках. В одних случаях информацию необходимо узнавать напрямую в различных отделах университета, в других – информация расположена на информационном ресурсе университета. При этом информационные ресурсы часто имеют сложную разветвлённую структуру, быстрый поиск в которой затруднителен, особенно для тех пользователей, которые работают с ресурсом впервые.

С другой стороны, в крупных учебных заведениях обучается большое количество человек. Например, по статистике общее количество учащихся в Московском государственном университете составляет 38150 человек, из которых 3907 человек – иностранные граждане. Посетителю учебного заведения может быть необходим один из его отделов, например – деканат, отдел кадров или библиотека. Однако, особенно при первом посещении учебного заведения, расположение необходимого отдела может быть ему неизвестно. В таком случае для того, чтобы найти необходимый отдел человек может воспользоваться указателями, помощью вахтера или встретившегося студента. Однако, часто в зданиях университета отсутствуют подробные указания о местоположении отделов, а информация от неподготовленных к решению данного вопроса людей

Статья получена 17 мая 2016.
Бородин В.А., АГТУ (e-mail:vitkt@yandex.ru).
Щербатов И.А., АГТУ (e-mail:sherbatov2004@mail.ru).

может оказаться недостоверной.

II. Описание системы

A. *Цели и задачи*

Целью данной работы является повышение эффективности взаимодействия между пользователем и интерфейсом консультирующей системы. Задачей является описание структуры и алгоритма работы подобной системы.

B. *Модель системы*

Для реализации системы консультирования с пониманием смысла запроса, а не только синтаксиса, необходимо понять смысл слов запроса. Также необходимо каким-то образом отобразить слова запроса к отделам учебного заведения, найти отдел учебного заведения максимально соответствующий запросу (рис.1).

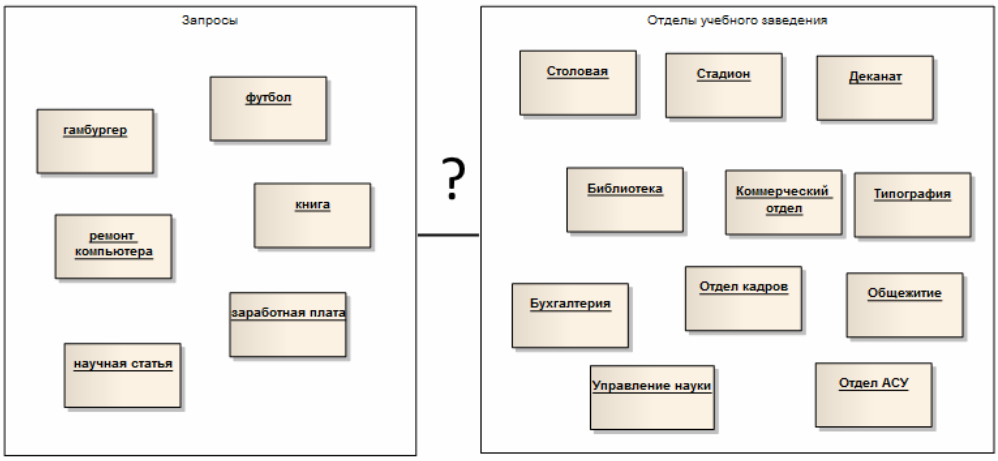

Рис. 1. Проблема реализации системы консультирования

Представим этот процесс в виде теоретико-множественной модели:

$$resultDep = nearsetDep(query, depSet), \quad (1)$$

где query – это строка запроса, depSet – множество отделов, resultDep – ближайший по смыслу отдел.

Каждому отделу университета, используемому в системе, был присвоен тег: слово, смысл которого ассоциируется с отделом. Например: спорткомплекс – «спорт», столовая – «еда», библиотека – «книга». От ситуации, отображённой на рисунке 1 мы перейдём к ситуации, изображённой на рисунке 2.

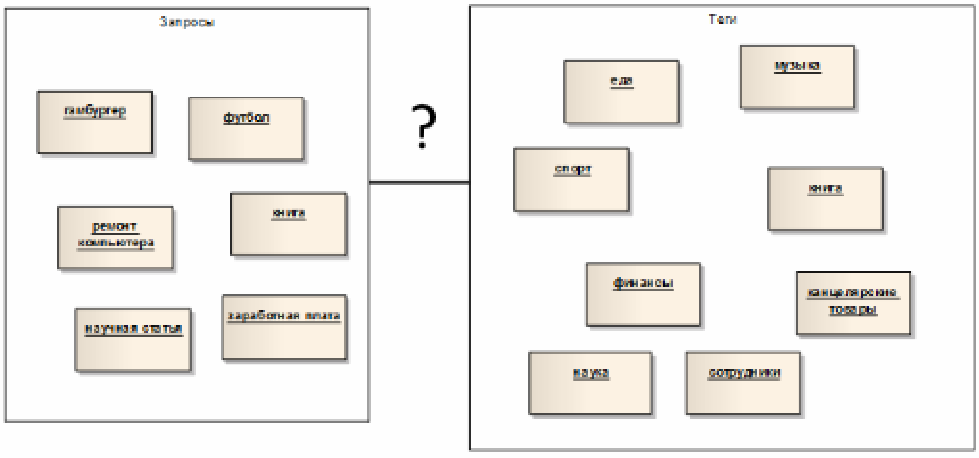

Рис 2. Использование тегов.

С точки зрения модели это будет представлено в виде множества тегов:

$$tagSet = getTagSet(depSet), \quad (2)$$

где tagSet – множество тегов, depSet – множество отделов, getTagSet – функция, присваивающая каждому отделу учебного заведения тег и формирующая из них множество.

Таким образом, вместо поиска максимально соответствующего отдела учебного заведения запросу мы ищем максимально близкое по значению слово:

$$resultDep = getDepOfTag(nearestWord(query, getTagSet(depSet))), \quad (3)$$

где query – запрос, depSet – множество отделов, getTagSet – функция, присваивающая каждому отделу учебного заведения тег и формирующая из них множество, getDepOfTag – функция, сопоставляющая тегу соответствующий ему отдел, nearestWord – функция, определяющая для запроса максимально близкое по смыслу слово из множества слов.

Реализация функции getDepOfTag и getTagSet не представляет сложной задачи, её можно решить, имея составленное администратором системы соответствие между тегами и отделами университета.

Для реализации функции nearestWord перейдём к термину «семантическая близость» или «семантическое подобие». Основные методы вычисления семантической близости можно разделить на две категории: топологические и статистические.

Топологические методы основываются на построенных семантических сетях между терминами. Статистические методы основываются на вычислении статистических метрик в корпусе документов. Для реализации алгоритма консультирования больше подходят топологические методы, так как для статистических методов необходим большой корпус документов, а одной из особенностей разрабатываемой системы является возможность поиска с использованием слов, не представленных на информационном ресурсе учебного заведения.

Топологические методы варьируются с простого подсчёта границ (Rada, 1989) до попыток вычисления коэффициентов на структуры сети, например, смены направлений связей (Hirst, St-Onge, 1998), относительной глубины (Sussna, 1993; Leacock, Chodorow, 1998) и плотности (Agirre and Rigau, 1996)[4]. Эти аналитические методы на данный момент конкурируют с методами статистики и машинного обучения. Помимо этого предложены гибридные методы, комбинирующие богатые знаниями источники, такие как тезаурус с менее богатыми – такими как статистика по корпусу документов (Resnik, 1995; Lin, 1998; Jiang, Conrath, 1997)[4].

Для определения семантической близости между запросом и тегом воспользуемся самым простым методом – подсчётом рёбер в семантической сети.

Тогда для поиска наиболее подходящего по смыслу тега необходимо выбрать тег с наименьшим расстоянием до слова запроса:

$$nearestWord(query, tagSet) = \min_{i=1..n}(dist(query, tagSet_i)) \quad (4)$$

На рисунке 3 представлена графическая модель информационной системы с использованием семантической сети, а также голосового ввода.

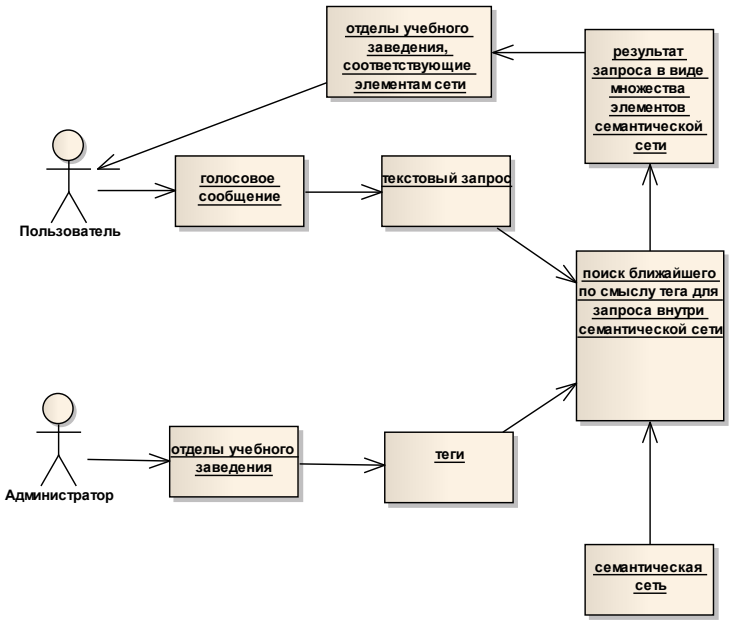

Рис. 3. Графическая модель системы

Администратор системы редактирует отделы учебного заведения и соответствующие им теги. Пользователь вводит голосовое сообщение, которое преобразуется в текстовый запрос, а затем разбивается на теги. Затем в семантической сети находятся ближайшие по смыслу теги, которым в соответствие ставятся отделы учебного заведения. Рассмотрим более подробно алгоритм обработки запросов

C. Алгоритм обработки запросов

В качестве примера опишем алгоритма запроса, содержащего только одно слово - одно слово - «карандаш». Во время инициализации программы система загружает семантическую сеть в оперативную память. Загруженные данные хранятся в виде списка слов и списка связей. Такое представление позволяет использовать алгоритмы над графами для данной системы. Помимо запроса в систему передаётся список тегов, соответствующих различным отделам учебного заведения, например – «книги», «письменные принадлежности», «музыка», «спорт» и др. С помощью данных индексных файлов слову из запроса, а также тегам ставятся в соответствие вершины графа – синсеты. В основе механизма классификации лежит алгоритм поиска в ширину. Поиск начинается с элемента,

содержащего слово из запроса. Поиск может считаться завершённым, если найден любой из тегов, ближайший в семантическом графе для этого слова. Таким образом, для слова «карандаш» первым найденным тегом будет являться тег «письменные принадлежности». Если слово имеет несколько значений, поиск выполняется над каждым из них, а результаты обработки объединяются. Опишем алгоритм работы системы пошагово:

- Шаг 1. Ввод запроса.
- Шаг 2. Загрузка списка тегов и отделов учебного заведения.
- Шаг 3. Загрузка семантической сети в виде ориентированного графа.
- Шаг 4. Получение списка вершин, соответствующих различным значениям слов из запроса.
- Шаг 5. Получение соответствующей вершины для каждого тега.
- Шаг 6. Запуск алгоритма поиска в ширину на графе от вершины одного из значений слова- запроса.
- Шаг 7. При достижении алгоритмом вершины, соответствующей какому-либо тегу, завершить поиск в ширину и перейти к шагу 9.
- Шаг 8. Если алгоритм поиска в ширину завершился, не достигнув какой-либо вершины- тега, перейти к шагу 10.
- Шаг 9. Добавить найденный тег в список результатов.
- Шаг 10. Если обработаны все значения слова-запроса, перейти к шагу 11. Иначе перейти к шагу 6, выполнив алгоритм для следующего значения слова-запроса.
- Шаг 11. Если не найдено ни одного тега, вывести сообщение об отсутствии результатов и завершить работу алгоритма.
- Шаг 12. Сопоставить каждому тегу отдел учебного заведения, исключить повторяющиеся результаты.
- Шаг 13. Вывести сопоставленные отделы учебного заведения, завершить работу алгоритма.

В семантической сети могут храниться не только отдельные слова, но и словосочетания. Поэтому для обработки подобных запросов используется специальный алгоритм: формируются различные комбинации сочетаний слов. Для каждого из этих сочетаний вычисляется результат классификации. Затем результаты ранжируются по числу слов в запросах-сочетаниях. Таким образом, результат обработки фразы, составленной из всего запроса, имеет большую релевантность, чем результат обработки составной части запроса. Например, запрос «финансовая помощь» система попытается обработать в виде одного словосочетания. Затем будут обработаны слова «финансовая» и «помощь» по отдельности. Тег, соответствующий полному запросу (например, «коммерческий отдел») будет находиться в верхней позиции в списке результатов, за ним будут следовать результаты обработки слов «финансовая» и «помощь». На последнем этапе работы программы найденным тегам ставятся в соответствие отделы учебного заведения. Если одному тегу соответствует несколько отделов учебного заведения, результаты объединяются, повторяющиеся элементы удаляются.

III. РЕАЛИЗАЦИЯ СИСТЕМЫ

В соответствии с данной структурой был разработан прототип системы. Разработанная система состоит из: подсистемы распознавания речи, подсистемы пользовательского интерфейса и подсистемы обработки запросов (рис. 4).

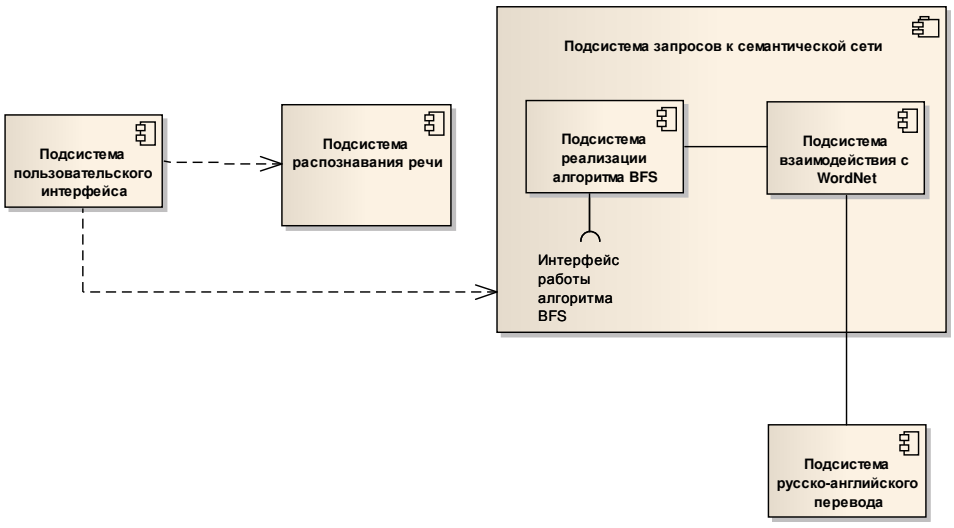

Рис. 4. Структура разработанного прототипа

Подсистема распознавания речи расширяет графический пользовательский интерфейс до естественного пользовательского интерфейса за счёт голосового взаимодействия. Система с подобным интерфейсом позволяет ускорить ввод текста на сенсорных экранах, а также улучшить интерактивность. В задачи подсистемы входят запись голосового сообщения и его распознание. Подсистема распознавания речи преобразует голосовое сообщение в текстовый запрос, который передаётся в подсистему взаимодействия с семантической сетью. Подсистема взаимодействия с семантической сетью выполняет следующие задачи: загрузку словаря WordNet, его преобразование в необходимый для реализации поисковых запросов виды, выполнение поисковых запросов, ранжирование результатов. Получив запрос из

подсистемы распознавания речи или из подсистемы графического интерфейса в текстовом виде, подсистема взаимодействия с семантической сетью выполняет обработку запроса.

База данных WordNet является базой данных для английского языка, однако в системе консультирования используются запросы на русском языке. Для решения данной задачи было выбрано использование системы перевода с русского языка на английский. Так как система консультирования предназначена для обработки не только слов, но и простых словосочетаний, для реализации перевода была выбрана система, позволяющая переводить не только отдельные слова, но и согласованные предложения. Для уменьшения затрат на разработку была выбрана клиент-серверная система, взаимодействие с сервером происходит путём вызова методов API.

Результат запроса передаётся обратно в подсистему графического интерфейса. Подсистема графического интерфейса обеспечивает взаимодействие пользователя с системой: содержит кнопку начала записи голосового сообщения, поле для текстового ввода сообщения и окно для вывода результатов запроса.

IV. Заключение

Таким образом, эффективным способом получения консультации об отделах учебного заведения может являться специально разработанная информационная система. Для повышения интерактивности консультирования система должна сочетать положительные качества поиска (возможность использования запросов) с удобством списков или интерактивных карт (отсутствие ввода через экранную клавиатуру). Также система должна понимать смысл запроса, а не оперировать только словами с похожим синтаксисом. Для решения данной проблемы может использоваться семантический поиск. Преимущества такого способа консультирования:

- Поиск на основе смысла слов, а не существующих документов
- Отсутствие необходимости в изучении иерархии интерфейса
- Возможность сочетания поиска с интерактивной картой или поисковой системой.

Развитие системы может предполагать обучение на основе выбранных пользователем вариантов – например, присваивание рёбрам графа весов и последующее их обновление. Также система может быть адаптирована для других предметных областей – например, для консультации пользователей в торговых центрах.

Библиография

[1] Басипов, А. А., Демич, О. В. Семантический поиск: проблемы и технологии. Компьютерное обеспечение и вычислительная техника. Вестник Астраханского государственного технического университета. Сер.: Управление, вычислительная техника и информатика — 2012. — № 1. - С. 104-111.

[2] Соболев М.С. Метод латентного семантического анализа: дис. магистра. МФТИ, Москва, 2007.

[3] Christiane Fellbaum. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). MIT Press, 1998 – 423 p.

[4] Alexander Budanitsky. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures / Graeme Hirst - Department of Computer Science University of Toronto Toronto, Ontario, Canada M5S 3G4.

[5] Eneko Agirre. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches / Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, Aitor Soroa - Institute of Formal and Applied Linguistics, Charles University in Prague.

[6] Alexander Budanitsky. Evaluating WordNet-based Measures of Lexical Semantic Relatedness/ Graeme Hirst - University of Toronto.

[7] Ted Pedersen. WordNet::Similarity / Siddharth Patwardhan, Jason Michelizz - Measuring the Relatedness of Concepts.

[8] 8. Jaap Kamps. Using WordNet to Measure Semantic Orientations of Adjectives / Maarten Marx and Robert J. Mokken and Maarten de Rijke - Language & Inference Technology Group, ILLC, University of Amsterdam.

Searching advising system for educational institution based on semantic network

Borodin V.A., Shcherbatov I.A.

***Annotation*— The article reviews a current trend analysis of search queries from semantics of expression. Authors use example of advising system for university`s visitors to describe model of such system based on semantic similarity. The paper consists description of algorithm for processing single word, phrases and prototype`s. This example shows the effectiveness of the proposed solutions.**

***Keywords* — semantic similarity, semantic network, advising system.**