

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Петрозаводский государственный университет»
Институт математики и информационных технологий
Кафедра прикладной математики и кибернетики

подпись соискателя

Бурдин Григорий Олегович

План-проспект выпускной квалификационной работе на степень магистра

**Разработка вопросно-ответной системы на основе латентно-
семантической поисковой модели**

09.04.02 – Информационные системы и технологии

Научный руководитель:

д-р техн. наук, доцент Л. В. Щеголева

подпись руководителя

Петрозаводск

2020

Аннотация

Целью данной работы является описание разработки вопросно-ответной системы на основе латентно-семантического поиска для конкретизированной предметной области – приёмная кампания Петрозаводского Государственного Университета.

Для достижения поставленной цели были решены следующие задачи: (1) сформирован корпус документов, содержащих информацию о предметной области, (2) сформирован поисковый индекс на основе корпуса документов, (3) разработан и спрограммирован алгоритм поиска вероятного ответа в документах корпуса по заданному пользователем запросу.

В рамках научно-исследовательской работы был проведен опрос с целью сбора основной информации о предметной области, а также типовых вопросов от предполагаемых пользователей. Был сформирован набор текстовых документов, содержащих основную информацию о различных аспектах обозначенной предметной области.

Результаты данной работы могут быть использованы для разработки более качественной системы. При наличии корпуса документов, удовлетворяющих требованиям разработанной системы, можно применять разработанный прототип, как рабочий вариант вопросно-ответной системы.

Введение

Вопросно-ответная форма получения информации является наиболее простой для человека. По этой причине за прошедшее десятилетие появилось множество систем, которые моделируют такую форму общения. Одним из участников в таком случае является программа или бот. На вход программе поступает вопрос пользователя, сформулированный на естественном языке. Вопрос обрабатывается семантическим анализатором, а на выходе ожидается ответ на данный вопрос, сформулированный также на естественном языке. Руководствуясь этим принципом, создаются чат боты и некоторые системы поддержки принятия решений. Качество ответов на вопросы в конкретной предметной области, получаемых от такой системы, может быть ниже, чем у специализированных экспертных систем. Данный фактор компенсируется тем, что вопросно-ответные системы являются широкопрофильными, то есть тематика задаваемых вопросов может быть из различных проблемных областей.



Рисунок 1. Схема системы

Для построения вопросно-ответных систем в основном используются три^[1] подхода:

1) Подход на основе информационного поиска

Данный подход основывается на поисковых алгоритмах, а именно поиске релевантных фрагментов в неструктурированных документах. Основной задачей при реализации данного подхода является качественный парсинг исходных документов для вычленения релевантной информации, а также формирование индекса документов, по которому будет производиться поиск.

2) Подход на основе знаний

В отличие от информационного поиска, данный подход предполагает наличие структурированной информации (база данных, хранение данных по определённым правилам и паттернам и т.п.). Также основополагающим

является положение о том, что вопрос содержит в себе часть ответа. Важным компонентом в вопросно-ответной системе, построенной по такому принципу, является семантический парсер.

3) Гибридный подход

Используются компоненты двух вышеописанных подходов.

Краткий обзор существующих исследований по теме

Основная концепция вопросно-ответных систем, разделения её на модули и некоторые методы решения задач, которые ставятся перед подомными системами, были описаны в публикации^[2].

Как было сказано выше, важным компонентом вопросно-ответной системы является семантический анализатор. В своей работе^[3] М.В.Мозговой рассматривает построение системы на основе семантического анализатора В.А.Тузова. Разрабатываемая автором система работает со строго заданной формой вопроса, содержащей вопросительное слово и некоторую часть ответа (что является важным для подхода на основе знаний).

Другой важный модуль вопросно-ответной системы, а именно модуль анализа вопроса, рассмотрели в своей публикации^[4] А.А.Соловьёв и О.В.Пескова. Особенность работы заключается в анализе структуры вопросов исключительно на английском языке.

Решить задачу поиска ответа на вопрос при помощи синтаксических деревьев и векторного описания вопросов, а также применением нейросетей, постарались авторы публикации^[5] из НИЯУ МИФИ. Разработанный алгоритм показывает хорошие результаты и обладает высокой точностью.

Тема пользуется большой популярностью для английского языка. Существует большое число публикаций, описывающих как готовые системы, так и отдельные модули. Так система AnswerBus^[6] принимает вопросы на различных языках (немецкий, английский, португальский, французский и некоторые другие) и возвращает ответ на английском. Система анализирует предложения, ранжируя потенциальные ответы и выбирая по определённому критерию наиболее релевантный. Рассматриваются^[7] также вопросно-ответные системы для не фактоидных вопросов, т.е. вопросов, на которые может выходить за рамки короткой заметки. Некоторые системы^[8] могут опираться не столько на лингвистические алгоритмы, сколько на избыточность данных для поиска и статистические вероятности правильности ответа.

Вопросно-ответные системы находят своё применение во многих прикладных отраслях. Так имеется много примеров разработок в области медицины^{[9][10]}, а также в мобильных системах^[11], где возможно расширение интерфейсов взаимодействия с пользователем.

План-проспект

В первом разделе описана процедура сбора информации о предметной области, а также процедура формирования корпуса документов, по которому должен производиться поиск, и структура этих документов. Был сформирован список тематик документов и список типовых запросов от пользователей, на основе которых в дальнейшем тестировалась система.

Во втором разделе описывается процедура создания поискового индекса для латентно-семантического поискового ядра, создание матрицы терм-документ и процедура вычисления и корректировки весов в данной матрице.

В третьем разделе описаны алгоритмы обработки запроса получаемого от пользователя, поиска документа, содержащего ответ и алгоритм поиска фрагмента документа, содержащего ответ на запрос пользователя. Также приведены результаты тестирования системы и вносимые на их основании изменения.

Заключение

По результатам работы была разработана вопросно-ответная система для ограниченной предметной области на основе информационного поиска. Индекс полученной системы может быть также применён для гибридной системы, так как часть документов соответствует требованиям, поставленным для системы на основе знаний. Применение поискового алгоритма без предобработки запроса дополнительными семантическими или синтаксическими анализаторами продемонстрировало, что такой подход может показывать неплохие результаты. Самый сложным и важным этапом разработки подобной системы является формирование индекса, потому что частота встречаемости термина зачастую не означает его важность.

Дальнейшим вектором разработки системы является улучшение алгоритма поиска, расширение индекса, а также создание модуля синтаксического анализа запроса для улучшения показателей точности и полноты поискового алгоритма.

Источники

- [1] Daniel Jurafsky, James H. Martin. Speech and Language Processing. Chapter 25. Question answering [Электронный ресурс] // URL: <https://web.stanford.edu/~jurafsky/slp3/> (дата обращения: 13.02.2020).
- [2] Allam A. M. N., Haggag M. H. The question answering systems: A survey //International Journal of Research and Reviews in Information Sciences (IJRRIS). – 2012. – Т. 2. – №. 3.
- [3] Мозговой М. В. Простая вопросно-ответная система на основе семантического анализатора русского языка //Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. – 2006. – №. 1.
- [4] Соловьёв А. А., Пескова О. В. Построение вопросно-ответной системы для русского языка: модуль анализа вопросов //Новые информационные технологии в автоматизированных системах. – 2010. – №. 13.
- [5] Наumenко А. М. и др. Разработка вопросно-ответной системы с нейросетевым обучением на базе современных свободных технологий //Иннов: электронный научный журнал. – 2017. – №. 2 (31).
- [6] Zheng Z. AnswerBus question answering system //Human Language Technology Conference (HLT 2002). – 2002. – Т. 27.
- [7] Oh J. et al. Non-factoid question-answering system and computer program : пат. 9697477 США. – 2017.
- [8] Brill E., Dumais S., Banko M. An analysis of the AskMSR question-answering system //Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. – Association for Computational Linguistics, 2002. – С. 257-264.
- [9] Abacha A. B., Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies //Information processing & management. – 2015. – Т. 51. – №. 5. – С. 570-594.
- [10] Cairns B. L. et al. The MiPACQ clinical question answering system //AMIA annual symposium proceedings. – American Medical Informatics Association, 2011. – Т. 2011. – С. 171.
- [11] Quarteroni S., Manandhar S. A chatbot-based interactive question answering system //Decalog 2007. – 2007. – Т. 83.