УДК 004.82

Автоматическое извлечение решеток понятий из текстов с помощью методов квантовой механики

Анастасия Сулягина
Университет ИТМО
Санкт-Петербург
anastasia.sulyagina@gmail.com

Научный руководитель: Бессмертный И.А
Университет ИТМО
Санкт-Петербург

Введение

Всю историю человечества знания аккумулировались в виде текстовых источников. Одной из важнейших задач информационного поиска является формализация этих знаний. Однако достаточно быстро обнаружилось, что формализация знаний требует слишком больших затрат человеческих и вычислительных ресурсов. Кроме того, даже высокая квалификация исполнителей не гарантирует точности, полноты и непротиворечивости оцифрованных знаний, поскольку сами эти знания не обладает данными свойствами. Тексты на естественном языке в основном базируются на человеческой логике, которая отличается от классической математической логики, что еще более усложняет извлечение знаний из этого вида данных и препятствует методам, основанным на математической логике, достигать качественных результатов.

Последние 30 лет проводятся обширные исследования квантовой логики [2] и ее применимости к прикладным задачам [4], [6]. В ходе исследований [3], [5] было обнаружено, что понятийный аппарат человека достаточно правдоподобно описывается эмпирическими законами квантовой механики, установленными для элементарных частиц. Можно предположить, что естественно-языковые тексты обладают теми же свойствами, что и понятийные модели в человеческом сознании. В следствие этого, перспективным направлением развития методов информационного поиска следует считать применение математического аппарата квантовой логики к естественно-языковым текстам.

Разработанный в этой статье алгоритм применяет методы квантовой логики к задаче извлечения иерархии понятий из текстов на естественном языке.

1. Цель работы

Целью данной работы является улучшение существующих методов извлечения понятий предметной области из естественно-языковых текстов с помощью нового алгоритма и создание удобной для индексирования и поиска решетки понятий, обладающей графовой структурой.

2. Базовые положения исследования

В данной работе было проведено исследование существующих подходов к извлечению понятий предметной области из текстов [1], [7], проанализированы их особенности и недостатки, и представлен новый метод.

Метод, использованный в этой работе, основан на применении теста Белла к документу и парам слов в нем. Теорема Белла [2] в квантовой физике иллюстрирует концепцию запутанности - когда две или более частиц в квантовом состоянии продолжают быть взаимосвязаными даже при большом физическом расстоянии. В применении к текстовым документам тест Белла может помочь определить, является ли пара слов взаимозависимой в данном контексте.

Для избавления от необходимости проверять тестом белла каждую пару слов в документе, для текста строится Hyperspace Analogue Language(HAL) [8] матрица, представление, хорошо подходящее для выделения словосочетаний, находящихся в одном контексте. Такие пары слов и будут проверяться на взаимосвязь тестом белла.

После вычисления теста Белла, для выделенных им взаимосвязанных пар слов текста подсчитывается метрика TF-IDF (Term Frequency - Inverse Document Frequency) [1] относительно контрастного корпуса документов. Она позволяет определить, являются ли выделенные термины специфичными для предметной области текста.

После выделения понятий предметной области для текстов корпуса, производится построение иерархий понятий для текстов

3. Результаты

В качестве эксперимента мы проанализировали корпус из 10000 контрастных документов [9]. В зависимости от предметной области и размера текста было выделено 20-200 концепций, образующих иерархию. Иерархии извлеченных понятий описывают предметную область текстов с достаточной точностью, позволяют определить тексты со схожей терминологией и могут использоваться для поиска информации или индексации.

Список литературы

[1] Daniel Jurafsky, James H. Martin (2009). Speech and Language Processing (2Nd Edition)

[2] Samson Abramsky, Ross Duncan (2006). A categorical quantum logic

[3] Diederik Aerts, Liane Gabora, Sandro Sozzo (2013). Concepts and their dynamics: A quantum-theoretic modeling of human thought

[4] Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, Keith van Rijsbergen. (2010). What can quantum theory bring to information retrieval?

[5] Alessandro Sordoni, Jian-Yun Nie, Yoshua Bengio (2013). Modeling term dependencies with quantum language models for IR

[6] Stephen Clark, Bob Coecke, Edward Grefenstette et al. (2013) A quantum teleportation inspired algorithm produces sentence meaning from word meaning and grammatical structure

[7] Bessmertny I.A (2018) Methods, models and software for building intelligent systems on a production knowledge model *ITMO University*, P. 304

[8] Kevin Lund, Curt Burgess (1996) Producing high-dimensional semantic spaces from lexical co-occurrence

[9] Wikipedia Monolingual Corpora *linguatools.org/tools/corpora/wikipedia-monolingual-corpora/*