УДК: 004.89
Название: Семантическое ранжирование документов на принципах квантовой теории
Автор: Семененко Евгений Константинович, Суров Илья Алексеевич, Университет ИТМО, Санкт-Петербург
Научный руководитель: Суров Илья Алексеевич, Университет ИТМО, Санкт-Петербург

Актуальность семантического поиска на основе квантовой теории

В настоящее время информационный поиск – это бурно развивающаяся область науки, популярность которой обусловлена экспоненциальным ростом объемов обрабатываемой информации, в частности в сети Интернет. Необходимость обработки и анализа таких данных представляет серьезную проблему для различных областей деятельности. Современные инструменты для анализа текстов и документов обладают рядом недостатков, устранение которых требует новых подходов к обработке и анализу информации. А именно, речь идет об учете «взаимодействия» пользователя с «умной» поисковой системой, что должно приводить к более точному формулированию контекста запросов и как следствие – релевантности выдаваемых документов. Одним из современных подходов, учитывающем контекстуальность при взаимодействии пользователя и «умной» информационной системы является квантовая когнитивистика, описывающая психологические аспекты принятия решений с помощью квантовых вероятностных методов и подходов квантовой теории измерений. В настоящее время наблюдается повышенный интерес к применению квантового формализма к задачам информационного поиска. Модели информационного поиска (логические, вероятностные и векторные) могут быть описаны с помощью квантового формализма Гильбертова пространства. При этом оказывается возможным учесть контекстуальность запросов.

Целью работы является предложить способ улучшения квантовоподобного алгоритма ранжирования при помощи теста Белла, учитывающего контекстуальность и совместимость различных запросов.

Общие сведения

Релевантность поисковой выдачи обусловлена информационным интересом пользователя и историей поисковых запросов. Необходимое обеспечение контекстуальной логики поисковых систем может быть выполнено на основе квантовой теории, которая изначально сконструирована для моделирования вероятностно-контекстуальных процессов микромира. В частности, квантово-теоретический параметр Белла [1], изначально разработанный для выявления контекстуальной связи удаленных физических процессов, может быть использован для измерения контекстуально-смысловой связи поисковых запросов.

Неравенство Белла (тест Белла) – математическое условие, которому обязана удовлетворять любая статистика дихотомических исходов двух (разнесенных в пространстве) измерений:

(1) $$S = |\langle AB\rangle - \langle AB'\rangle + \langle A'B\rangle + \langle A'B'\rangle)| \leq 2,$$

где A и B – подсистемы, над которыми проводятся эксперименты, имеющие каждый по два возможных исхода, кодируемые значениями ±1. Примером из физики может послужить система из пары электронов, над каждым из которых проводится эксперимент по измерению спина. Нарушение неравенства Белла, когда значение параметра S в диапазоне от 2 до $2\sqrt{2}\approx 2.82$, составляет доказательство контекстуальной взаимообусловленности для совокупности систем и экспериментальных процедур.

В оригинальном эксперименте [2] наблюдаемые A и B ставятся в соответствие в части поискового запроса, представляемые HAL-векторами [6] контекста в словарном пространстве документа. В данной работе удалось установить связь между ранжированием трех документов по релевантности и длиной контекстуальной связи, необходимой для нарушения неравенства Белла.

Результаты

Мы предлагаем улучшить алгоритм ранжирования [2], представляя части поискового запроса векторами не в исходном словарном пространстве HAL, а в семантическом пространстве меньшей размерности. Это представление строится методом латентного семантического анализа (LSA [3]), использующего сингулярное разложение матрицы словарных контекстов HAL. Мы ожидаем, что такой подход позволит учесть ассоциативно-смысловое содержание текста и поискового запроса. На данный момент получены следующие промежуточные результаты:

1. Эксперимент [2] воспроизведен для русского языка. Выполнено предварительное ранжирование текстов статей из Википедии релевантных тематике поискового запроса «Языки программирования».
2. Разработана модификация алгоритма [2] ранжирования на основе теста Белла над векторами документа и запроса в семантическом пространстве. Выполнен предварительный анализ результатов ранжирования. В соответсвии с ожиданием нарушения неравенства Белла происходит при меньшем размере контекста, что указывает на более плотное предствление смысла в семантическом пространстве пониженной размерности.
3. Для дальнейшего тестирования предложенного и перспективных алгоритмов ранжирования поисковой выдачи на квантовых принципах создана разметка релевантности выборки из 40 тыс. документов и 1,5 тыс. запросов на основе экспертных оценок.

Литература

1. Bell J.S. Speakable and Unspeakable in Quantum Mechanics. Cambridge University Press, 1993.
2. Barros J. и др. Contextual Query Using Bell Tests // International Symposium on Quantum Interaction. 2014. С. 110–121.
3. Deerwester S. и др. Indexing by Latent Semantic Analysis // Journal of the American Society for Information Science. 1990. Т. 41, № 6. С. 391–407.
4. Rijsbergen C.J. van. The Geometry of Information Retrieval. Cambridge University Press, 2004.
5. Li J. и др. An adaptive contextual quantum language model // Phys. A Stat. Mech. its Appl. 2016. Т. 456. С. 51–67.
6. Lund K., Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence // Behavior Research Methods, Instruments, & Computers. 1996. 28 (2). 203-208.

Автор: _____/Семененко Е.К.
semenenko.e.k@yandex.ru

Научный руководитель, автор: _____/Суров И.А.

Руководитель образовательной программы: _____/Алоджанц А.П.