УДК 519.767.6

ЭФФЕКТЫ НЕСЕПЕРАБАЛЬНОСТИ В КОГНИТИВНОМ СЕМАНТИЧЕСКОМ ПОИСКЕ ИНФОРМАЦИИ

А.П. Алоджанц
Университет ИТМО
Россия, 197101, г. Санкт-Петербург, Кронверкский проспект, д.49
E-mail: alexander_ap@list.ru

А.В. Платонов
Университет ИТМО
Россия, 197101, г. Санкт-Петербург, Кронверкский проспект, д.49
E-mail: aplatonovv@gmail.com

И.А. Бессмертный
Университет ИТМО
Россия, 197101, г. Санкт-Петербург, Кронверкский проспект, д.49
E-mail: igor_bessmertny@hotmail.com

Е.К. Семененко
Университет ИТМО
Россия, 197101, г. Санкт-Петербург, Кронверкский проспект, д.49
E-mail: semenenko.e.k@yandex.ru

Ключевые слова: информационные поисковые системы, принятие решений, квантовая когнитивистика, квантовое запутывание, контекстуальность, машинное обучение.

Аннотация: В работе исследован аналог тест Белла из квантовой физики, позволяющий определить наличие несепарабельности квантовых состояний применительно к задаче семантического поиска информации и ранжирования документов. В численном эксперименте за основу взят алгоритм HAL, позволяющий получить векторное представление слов (в Гильбертовом пространстве) на базе словарного индекса и учитывающий порядок слов в документе. Показано, что между двумя словами из запроса пользователя существуют определенные (не учитываемые при классическом вероятностном описании) зависимости (контекст) с точки зрения подпространства, генерируемого документом поисковой выдачи. Под подпространством подразумевается пространство, генерируемое вектором документа, полученным в результате объединения векторов слов, составляющих его. Предсказано, что выявленная таким образом контекстуальность может проявляться на когнитивном уровне (мышлении) человека как при формировании тех или иных текстов, так и запросов к ним.

1. Введение

В течение последнего десятилетия стремительный рост информационных ресурсов в части обмена и обработки информацией привел, по сути, к экспоненциальному росту объемов обрабатываемых данных, большую часть которых составляют плохоструктурированные (или вообще не структурированные) данные. Необходимость обработки и анализа таких данных в режиме реального времени представляет серьезную

проблему в различных областях экономики, финансов, социальной сфере, имеет непосредственное отношение к безопасности общества.

Тематическое моделирование, как одно из приложений машинного обучения является важным инструментом для анализа современных текстов и документов и имеет непосредственное применение в задачах информационного поиска. Несмотря на определенные преимущества, оно обладает 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. В этой связи модель семантического пространства, основанная на так называемом подходе мешка слов (*bag-of-word*), не учитывающем порядок слов, когда смысл кодируется счетчиками слов, которые входят в контекст объекта, выглядит весьма ограниченной.

Одним из современных подходов, учитывающем контекстуальность при взаимодействии пользователя и «умной» информационной системы является *квантовая когнитивистика* [1]. Интересно заметить, что первые попытки описания психологических аспектов в принятии решений с помощью квантовых вероятностных методов и подходов квантовой теории измерений были предприняты достаточно давно, еще в СССР [2]. В настоящее время наблюдается повышенный интерес к применению квантового формализма к задачам информационного поиска, см. напр., [3-6]. В частности, показано, что модели информационного поиска (логические, вероятностные и векторные) могут быть описаны с помощью квантового формализма Гильбертова пространства. При этом оказывается возможным учесть контекстуальность запросов [4]. В [5] сформулирован принцип ранжирования квантовой вероятности (Quantum Probability Ranking Principle), являющийся обобщением известного принципа ранжирования вероятности, используемого для оценки критериев ранжирования выдаваемых документов, и учитывающего зависимости между документами. При этом для того, чтобы получить наилучшую общую эффективность поиска, информационная система ранжирует документы не только в порядке убывания вероятности их релевантности для пользователя, нуждающегося в информации, но и учитывает эффекты «квантовой интерференции». Как показано в [6], уместными являются здесь аналогии с квантовой оптикой.

Целью данной работы является демонстрация квантовоподобного алгоритма ранжирования при помощи теста Белла с учетом вопросов контекстуальности и совместимости различных запросов.

2. Тест Белла в задаче когнитивного семантического поиска информации

2.1. Неравенство Белла и его интерпретация

В рассматриваемой задаче наиболее адекватен инструментарий квантовой теории, которая изначально сконструирована для моделирования несовместимых физических экспериментов – ситуаций, когда конкретная экспериментальная конфигурация позволяет согласованно определить некоторый набор физических величин, которые, однако, становятся неопределёнными при переходе к другой конфигурации эксперимента. В квантовой когнитивистике, такие несовместимые экспериментальные ситуации соответствуют несовместимым когнитивным контекстам, использование которых при принятии решений ведёт к нарушениям классической (булевской) логики. Следствием квантовой теории является возможность корреляций результатов измерений (проводи-

мых в физике над удалёнными системами), более сильных, чем это допускается классическими теориями со скрытыми параметрами [7].

Пусть имеются две подсистемы A и B, и проводятся эксперименты соответственно, имеющие каждый по два возможных исхода, кодируемые значениями ± 1. этому условию удовлетворяет, например, система из пары электронов, над каждым из которых проводится эксперимент по измерению спина. Каждый из экспериментов A и B проводится в двух вариантах: A и A', B и B', которые в рассматриваемом примере различаются направлением спинового измерения: 0°, 90°, 45° и 135° соответственно. Эти направления комбинируются четырьмя возможными способами: $\{A, B\}$, $\{A, B'\}$, $\{A', B\}$, $\{A', B'\}$, причём в каждой конфигурации для получения статистически-значимого набора (вероятностных) исходов, эксперимент проводится многократно. Для выявления таких корреляций в квантовой теории служит т.н. *неравенство Белла*

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

математическое условие, которому обязана удовлетворять любая статистика дихотомических исходов двух (разнесенных в пространстве) измерений. Неравенство (1) выполняется для сепарабельных состояний, когда справедливо условие $\langle \dots \rangle = \langle \dots \rangle\langle \dots \rangle$ факторизации для средних. В квантовой теории теорема Цирельсона предсказывает существование таких квантовых состояний двух подсистем, для которых определённые измерения над ними могут давать значение параметра S в диапазоне от 2 до $2\sqrt{2} \approx 2.82$. В физике эти состояния называют запутанными (entangled), или, несепарабельными [7]. Нарушение неравенства Белла составляет доказательство контекстуальной взаимообусловленности для данной совокупности систем и экспериментальных процедур.

2.2. Алгоритм теста Белла

В работе нами предложен эксперимент (тест Белла) с проверкой соответствующих неравенств, который можно провести с объектами в семантическом Гильбертовом пространстве. В эксперименте в качестве алгоритма для получения векторного представления слова использовался алгоритм *HAL* (*Hyperspace Analogue to Language*), который позволяет, в отличие от популярного «мешка слов» учесть порядок слов в предложении и тем самым повысить точность определения зависимостей между словами.

Алгоритм HAL позволяет получить векторное представление слов на базе словарного индекса (соответствия слова некоторому уникальному числовому идентификатору) и обрабатываемого набора документов. Для получения такого представления строится матрица в ячейках которой содержится сумма расстояний между словом, соответствующим строке матрицы и словом, соответствующим столбцу матрицы в рамках корпуса текстов. При этом моделируется отношение “расстояние от слова в строке до слова в столбце, если слово в столбце стоит справа”. Таким образом, берется в расчет порядок слов в предложении. Более того, расстояние рассчитывается не между всеми словами в матрице, а только между теми парами слов, которые находятся на расстоянии, не превышающем заранее определенное расстояние, называемое размером окна HAL.

Для проведения эксперимента использовались текстовые файлы, полученные из нескольких статей Википедии. В частности, были выбраны статьи по тематике «Язык программирования» (статья «Язык программирования», «Программирование», “Java”, “C++”) и содержимое этих статей, за вычетом верстки было подвергнуто препроцессингу. В процессе чтения файлов, строится текстовый индекс, который содержит нормализованные формы слов и их соответствие уникальному числовому идентификатору. Данный числовой идентификатор используется для определения одной координаты вектора в векторном пространстве слов, полученных алгоритмом HAL. После получения индекса слов для обрабатываемого документа строится матрица HAL (см. следую-

щий подраздел), которая позволяет получить векторное представление слов документа, а среднее значение суммы этих векторов - вектор документа. Векторы слов запроса получаются по той же матрице HAL. Полученные векторы документа и запроса используются для расчета параметра Белла.

Таким образом, по мере получения индекса и векторных представлений слов, слова пользовательского запроса (в данном эксперименте, например, это был «Язык программирования») приводятся к векторной форме, которая нормализуется, и на базе этих слов строятся два базиса, задающие подпространства векторов, соответствующих двум словам запроса. Для получения базиса используется алгоритм ортогонализации Шмидта. Вектор документа раскладывается по этим базисам:

(2) $$\left|D_{w_1}\right\rangle = a|+\rangle_A + b|-\rangle_A;\ \left|D_{w_2}\right\rangle = c|+\rangle_B + d|-\rangle_B,$$

где $\left|D_{w_1}\right\rangle$ – вектор документа; $|+\rangle_A$ ($|-\rangle_A$) и $|+\rangle_B$ ($|-\rangle_B$) – базисы, в которых запросы A и B полностью релевантны (не релевантны) документу. Далее, определяем (проективные) операторы измерений для запросов A и B, соответственно:

(3) $$A_x = |+\rangle_{A\ A}\langle -| + |-\rangle_{A\ A}\langle +|;\ B_x = |+\rangle_{B\ B}\langle -| + |-\rangle_{B\ B}\langle +|;$$
$$A_z = |+\rangle_{A\ A}\langle +| - |-\rangle_{A\ A}\langle -|;\ \ B_z = |+\rangle_{B\ B}\langle +| - |-\rangle_{B\ B}\langle -|.$$

Операторы, относящиеся к одним и тем же запросам (например, к A или, к B) некоммутируют между собой и соответствуют операторам измерения проекций спина в квантовой физике. Оператор A_z позволяет определить степень соответствия документа первому слову запроса. При этом, если вектор документа в базисе первого слова представлен вектором $[a, b]$ (2), то вероятность того, что документ относится к тематике первого слова может быть получена по правилу Борна $\langle A_z\rangle_D = \langle D|A_z\ |D\rangle = a^2 - b^2$.

Для измерений (запросов) используем также комбинации $B_+ = (B_z - B_x)/\sqrt{2}, B_- = -(B_z + B_x)/\sqrt{2}$, определяющие степень соответствия документа второму слову в повернутом базисе. Заметим, что расчеты можно проводить в базисе одного из слов (например, A), Полученный набор операторов используется для вычисления параметра Белла запросов (ср. с (1)):

(5) $$S_q = |\langle A_z B_+\rangle + \langle A_x B_+\rangle| + |\langle A_z B_-\rangle - \langle A_x B_-\rangle|.$$

3. Результаты

Результаты теста Белла для текстов на русском языке по тематике «язык (слово A) программирования (слово B)», взятых из Википедии, показаны на рис.1. Видно, что тест Белла существенно зависит от размера окна HAL при моделировании семантического пространства; в пределе все четыре графика стремятся к значению квантового предела $2\sqrt{2}$. Это может быть объяснено тем, что с увеличением размера окна при построении HAL-матрицы, термины, входящие в запрос пользователя, от которого считался тест Белла, при больших размерах окон чаще пересекаются контекстами друг с другом. То же самое можно сказать и о значениях, меньших 2. Однако ,принципиально, что все кривые на рис. 1 достигают зоны «квантовой запутанности» в разные моменты. Так, раньше всех этой точки достигает документ «Язык программирования», который непосредственно относится к тематике запроса. Следом идут документы ”Java”, ”C++” и в конце – «Программирование». Можно сделать вывод ,что такой тест может быть использован как признак релевантности (контекстуальности) статьи к тематике запроса.

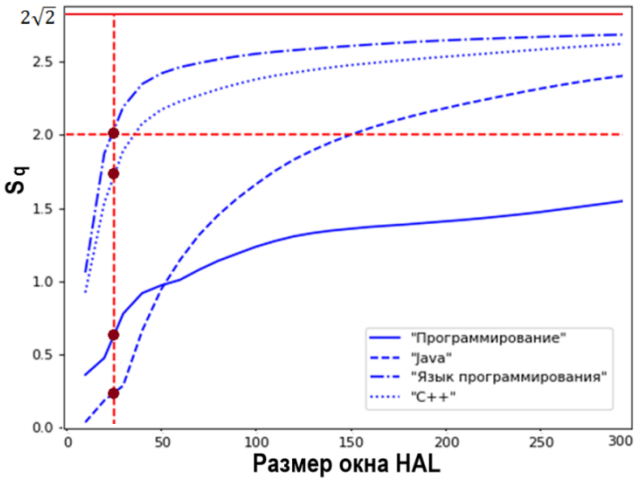

Рис. 1. График зависимости результата по тесту Белла четырех документов от размера окна при построении HAL-матрицы. Точками обозначены значения параметра Белла, указывающие на ранжирование соответствующих документов при запросе.

Таким образом, большие значения по тесту Белла не означают большую запутанность, однако признаком присутствия заданной тематики в статье может служить размер окна при вычислении HAL-матрицы, на котором могут проявляться признаки «квантовой запутанности», или, несепарабельности.

Более того, данный тест может быть использован для разделения текстовых документов по тематике запроса и документы, более близкие к тематике запроса могут показывать значения теста больше 2 на окнах меньшей размерности. Данное интересное свойство может быть использовано в нескольких случаях.

Во-первых, документы можно ранжировать по размеру HAL-окна, при которых достигается значение большее 2. Более релевантные документы будут иметь меньшие размеры окна.

Во-вторых, данный тест можно использовать для извлечения терминов предметной области: выбрав окно фиксированного размера можно извлекать парами цепочки слов, для которых тест Белла демонстрирует значения выше пороговых. В дальнейшем планируется использовать этот тест для более масштабной проверки этих гипотез.

Список литературы

1. Busemeyer J.R., Bruza P.D. Quantum models of cognition and decision. Cambridge University Press, 2012. 408 p.
2. Гуревич И.И., Фейгенберг И.М. Какие вероятности работают в психологии? Вероятностное прогнозирование деятельности человека. М.: Наука, 1977. С. 9-21.
3. Van Rijsbergen C.J. The Geometry of Information Retrieval. Cambridge University Press: New York, USA, 2004. 150 p.
4. Piwowarski B., Frommholz I., Lalmas M., van Rijsbergen C.J. What can Quantum Theory Bring to Information Retrieval // Proceedings of the CIKM 2010, Toronto, ON, Canada, 26-30 October 2010. P. 59-68.
5. Zuccon G., Azzopardi L. Using the Quantum Probability Ranking Principle to Rank Interdependent Documents // Proceedings of the ECIR 2010, Milton Keynes, UK, 28-31 March 2010. P. 357-369.
6. Zhang P., Song D., Zhao X., Hou, Y. Investigating Query-Drift Problem from a Novel Perspective of Photon Polarization // Proceedings of the ICTIR 2011, Bertinoro, Italy, 12-14 September 2011. P. 332-336.
7. Peres A. Quantum theory: concepts and methods. Dordrecht-Boston-London, Moscow: Kluwer Academic Publishers, 2002. 464 p.