УДК 004.623

USING BELL TEST FOR REALISING A QUANTUM-LIKE SEMANTIC MODEL FOR TEXT RETRIEVAL IN ARABIC TEXTS

Шакер Алаа (Национальный исследовательский университет ИТМО)
Научный руководитель – проф. ФПИиКТ, Бессмертный И.А.
(Национальный исследовательский университет ИТМО)

In this work, we try to release a new text retrieval algorithm from Arabic texts; the new algorithm considers the relation between words in the text instead of plain statistical analysis.

Introduction. With the steady increasing of information resources, the importance of information retrieval algorithms increases that try to reduce access time to information; traditional algorithms use keywords or some statistical methods like TF-IDF for finding target texts.

However, that ignores an important point; the meaning of text, the meaning of text is created gradually when person read each word, the meaning is changed by adding new information. In other words, the relation between words of text creates the meaning of text.

Our algorithm tries to detect the entanglement between two words in text by applying Bell test, which shows the analogy between natural language and quantum-like systems.

The main part. The algorithm consists of six steps: first getting the query and text, then the step of processing text and query (get the infinitive of each verb, root of each adjectives, and the singular status of each noun).

Third step is to build the Hyperspace Analogue to Language (HAL) matrix; here we want to pay attention to two points that characterize the matrix HAL:

- HAL is a sensitive array for the links between two terms: row and column vectors record the co-occurrence information of previous and later words separately.
- The value of HAL vectors is affected by the size of window; a wider window means greater chance for associations between two terms, but the large size may suffers from underfitting. On other side, the small window size mean a strong association between two terms but also may suffers from overfitting.

Fourth step, extracting three vectors from HAL matrix, the first is D vector "vector represents the whole text" as the normalized sum of all the word-vectors of the matrix of the document, the remaining two vectors, Dw1 and Dw2, represent pair of the query words.

Fifth step is to get orthogonal bases from two vectors {Dw1,Dw2}, For getting these bases, we use the rule called "Gramm-Schmidt", by applying this converting rule on {Dw1, Dw2} we get the first orthogonal bases $\{\ u_1,\ u_2\}$.By applying Gramm-Schmidt on {Dw2,Dw1}we get the second orthogonal bases$\{\ v_1, v_2\}$, after that we can project the vector of document D on these two bases $\{\ u_1, u_2\}$,$\{\ v_1, v_2\}$.

Last step is to apply Bell test inequity (CHSH) for detecting if there is either entanglement between these two words in text or not.

If the result is between $\{2\ ,\ 2*\sqrt{2}\}$ that means the entanglement between the two words of query in the text , and that means this text is relevant to subject of user search. If Bell's test value under 2 that means there is not entanglement, and this text is not relevant to the subject search.

Conclusion. The algorithm can be used for text retrieval, and the results of algorithm sometimes are better than traditional algorithms, it takes care important thing, which is the entanglement between words in the text, that is considered new method in text retrieval.

This algorithm can also be used for classification the text into classes or categories.

We can use it in ordering the result of search process, because the good search engine that return the texts are relevant to subject of query on the top of un-relevant texts, by using value of Bell's test.