Reinforcement Learning-driven Information Seeking: A Quantum Probabilistic Approach

Amit Kumar Jaiswal[0000−0001−8848−7041], Haiming Liu[0000−0002−0390−3657], and Ingo Frommholz[0000−0002−5622−5132]

University of Bedfordshire
Luton, United Kingdom
{amitkumar.jaiswal,haiming.liu,ingo.frommholz}@beds.ac.uk

Abstract. Understanding an information forager's actions during interaction is very important for the study of interactive information retrieval. Although information spread in an uncertain information space is substantially complex due to the high entanglement of users interacting with information objects (text, image, etc.). However, an information forager, in general, accompanies a piece of information (information diet) while searching (or foraging) alternative contents, typically subject to decisive uncertainty. Such types of uncertainty are analogous to measurements in quantum mechanics which follow the uncertainty principle. In this paper, we discuss information seeking as a reinforcement learning task. We then present a reinforcement learning-based framework to model the foragers exploration that treats the information forager as an agent to guide their behaviour. Also, our framework incorporates the inherent uncertainty of the foragers' action using the mathematical formalism of quantum mechanics.

Keywords: Information Seeking · Reinforcement Learning · Information Foraging · Quantum Probabilities.

1 Introduction

Web searchers, in general, move from one webpage to another by following links or cues while keeping the consumed information (intake) with itself without attaining a generalised appetite (information diet) in possession of uncertain and dynamic information environments [1, 2]. In general, the evolution of information patterns from user interaction keeps searchers in an information seeking process to not consume optimised information diet (the information goal). So, there needs to be a mechanism that can guide the foragers during their search process in order to set a realistic information appetite. User interaction is an important part of the search process which can enhance the search performance and the information foragers' search experiences and satisfaction [3, 5, 15]. User

action and their dynamics during search play an important role in changing behaviour and user belief states [12]. It has been recently demonstrated that action behaviour representations can be learned using reinforcement learning (RL) [11] by extrapolating a policy in two components - action representation and its transformation. To effectuate the information foragers' (or searchers'/users') [6] cognitive ability during the search, we treat the searcher as an RL agent which follows Information Foraging Theory (IFT) [4], to understand how the users can learn in an ongoing process of finding information. Furthermore, the learning ability of the users can be signalled by the RL approach through giving a free-choice of search scenarios in an uncertain environment. For instance, the information seeker must optimise the trade-off between exploration by sustained steps in the search space on the one hand and exploitation using the resources encountered on the other hand. We believe that this trade-off characterises how a user deals with uncertainty and its two aspects – risk and ambiguity during the search process [6]. Therefore the pattern of behaviour in IFT is mostly sequential. Risk and ambiguity minimisation cannot happen simultaneously, which leads to an underlying limit on how good such a trade-off can be. This lets the information foraging perspective of information seeking converge with the developing field of quantum theory [6]. Moreover, web search engines enable their users to efficiently access a large amount of information on the Web, which in turn leads search users to learn new knowledge and skills during their search processes. When the users search to obtain knowledge, their information needs[1] are mostly varied, open-ended, and seldom not clear at the start. Such types of search sessions generally span multiple queries and involve rich interactions, therefore our aim is to model such kind of information foraging process where the users' cognitive state changes during search.

Due to its inherently complex and intense interactive nature, the effective and interactive information foraging process is exigent for both the users and the search systems. Hence, our focus is to incorporate contextual semantic information in modelling the information forager with the usage of the mathematical framework of quantum theory, i.e. quantum probabilities based on geometry. Specifically, we propose a quantum-inspired reinforcement learning approach that (a) models the information foragers' behaviour, where action-selection (or policy) is leveraged as an Actor-critic method [22] to enhance the agent's experience in a text query-matching task; (b) learns the policy where query representation is parameterised using quantum language models, with a focus on the interaction across multi-meaning words.

2 Related Work

This section covers aspects of reinforcement learning, Quantum theory in dynamic information retrieval (IR), in particular, interactive information retrieval [8, 7], and Information Foraging theory.

[1] we consider an IN is expressed by a query or series of queries

Reinforcement Learning in Information Retrieval: Humans' transfer of information to other animals is a common method of learning and interaction, which is generally called reinforcement learning. Reinforcement learning [10] (RL) techniques are motivated by our sense of decision making in humans which appears to be biologically rooted. Within such biological roots [9], when an information foragers' action ends up with a disadvantageous consequence (or negative payoff), such action will not be counted in the future; whereas, if his/her action leads to a successful consequence (or positive reward), it will happen again. User involvement in information searching is primarily a decision making (or action taking) process [20], where users reflect identical RL features during this process. We will adopt RL models to manifest the mechanisms prevailing users' learning of information from searching. Previous work [21] found that a search system's information can be enriched to advance search intention and automate the difficult query reformulation by modelling the search context. Reinforcement learning is an important method that can let the system employ the search context and relevance feedback simultaneously. Also, this approach allows the system to deal with exploration (widening the search among different topics) and exploitation (moving deeper into generic subtopics) which has been supportive in information retrieval [23, 24]. Exploration and exploitation methods are usually employed in tasks associated with recommender systems or information retrieval, such as foraging strategies [25], recommendation [26] or image retrieval [27]. However, reinforcement learning is mainly used by search/retrieval systems [31], which collect users' interests and habits over a continuous period, while in a specific search scenario the users in a given search session are more interested in the holistic improvement of the search results than relying on arbitrary future search sessions.

Quantum Theory and Information Retrieval: Quantum Theory (QT) has been matured to reinforce the search potential by employing the mathematical formalism of quantum mechanics to information retrieval [28]. The aim of introducing the QT formalism was to elucidate the implausible behaviour of micro-level search actions, which classical probability theory may not be able to model. Furthermore, it is an expressive formalism that can combine prominent probabilistic, geometric and logic-based IR approaches. The mathematical foundation of the Hilbert space formalism was introduced in [29] to apply this mathematical framework outside of Physics. We refer to events as a subset of a sample space of all potential events in classical probability theory, whereas in Quantum theory, the probabilistic space is geometrically defined, and the representation of it becomes an infinite set of angles and distance commonly named as an abstract vector space — or, more appropriately, a finite or infinite-dimensional Hilbert Space denoted by $\mathcal{H}$. Each and every event is depicted as a subspace of the Hilbert Space. To represent the n-dimensional vectors that compose a Hilbert Space, the Dirac notation is widely adopted, using *ket* and *bra* nomenclatures. More concretely, this means representing one given vector ψ as $|\psi\rangle$ and its transposed view, ψ^{T} as $\langle\psi|$. Also, the vectors under consideration in a Hilbert space are usually unit vectors (their length is 1). A projection onto a subspace induced

by a vector $|\psi\rangle$ is denoted by the operation resulting in a matrix $|\psi\rangle\langle\psi|$. In this subspace, the vectors contained within are again normalised[2], and the projection of events represented as vectors, again, is performed by the $|\psi\rangle\langle\psi|$ operation. Unit vectors interpreted as state vectors induce a probability distribution over events (subspaces), and the product resulting from the mentioned operation is called *density matrix*. We use so-called observables to perform a measurement of the outcomes (which are eigenvalues).

The major similarity between quantum mechanics (QM) and information retrieval (IR) is understanding the interaction between a user (the observer in QM) and the information object under observation [28]. The core connection between QM and IR stems from the probabilistic features, where there is an observation of agreement for the preface of conditional probabilities allied with interference effects dominating to some contextual measure (cognitive, subjective character) when consolidating varied objects[3]. In QT, we can represent user information needs with state vectors, and the query/observable, eigenvalues and the probability of obtaining single eigenvalues or objects as a measure of the degree of relevance to a query [8]. Earlier QM was incorporated withing the RL algorithmic approach to generalise on filtering favourable user actions [30].

Information Foraging Theory: Information Foraging theory (IFT) [4] was developed to understand human cognition and their behaviours. IFT provides stipulated constructs adopted from optimal foraging theory which includes predators conforming to humans who seek for information (or prey). It has three constructs, one of which delineates searches (or Search engine result page (SERP)s) in the user interface sections, referred to as *information patches*; *information scent* helps users make use of perceptual cues, such as web links spanning small snippets of graphics and text, consecutively to make their navigation decisions in selecting a specific link. The purpose of such cues is to characterise the contents that will be envisaged by trailing the links. Finally, *information diet* allows users to narrow or expand diversities of information sources based on their profitabilities (appetite).

Information Foraging is an active area of IR and information seeking due to its sound theoretical basis to explain the characteristics of user behaviour. IFT has been applied to model users' information needs and their actions using information scent [13]. However, it has been previously found that information scent can analyse and predict the usability of a website by determining the website's scent [14]. Liu et al. [15] demonstrated an IFT-inspired user classification model for a content-based image retrieval system to understand the users' search preference and behaviours by functioning the model on a wide range of interaction features collected from the screen capture of different users' search processes. Recent work [16, 17] studied the effects of foraging in personalised recommendation systems by inspecting the visual attention mechanism to understand how

[2] there may be some vectors which are not necessarily normalised

[3] https://www.newscientist.com/article/mg21128285-900-quantum-minds-why-we-think-like-quarks/

users follow recommended items. Such user-item interactions can also be seen in query-level interactions i.e., in query reformulation scenarios where IFT- and RL-like models [18, 19] provide better explainability.

3 Information Seeking As Reinforcement Learning Task

A searcher during the search process has to investigate several actions before selecting any of it, with unknown reward. They explore each result back and forth to estimate the optimal patch based on the reward. This scenario of information seeking can be interpreted as a reinforcement learning task where the search process, involving an agent to interact with the search environment, is cost-driven. Assessing positively rewarded actions (from searcher's incurred costs) by the agent within an uncertain environment can potentially optimise the foragers' choice in finding the information. From an IFT perspective, positively rewarded actions can be drawn as exploitation whereas the available actions as exploration provided the information must be scattered between patchy environment. The fundamental aspect of reinforcement learning is to "learn by doing with delayed reward", which emerges as a major connection to information seeking (especially user interaction in IR and recommendation tasks) and it also interprets the foraging process of a searcher. The seeker's goal is to quickly locate a relevant patch (document, image, etc). However, the information seeker has no prior knowledge of the rewards from assessed patches and they keep exploring each of it. The seeker interacts with the search system to explore which results in relevant information elicits the rewards distribution (information scent patterns) between information patches; often the access to patches with minimum reward can signify an optimal patch that the seeker has spent less time on for exploitation. An information seeker spending less time assessing each information patch leads to partially-relevant information about the seeking process that elicits the rewards distribution between the patches, and it gives rise to exploitation of a patch with less than the optimal rewards. Hence, the longer a seeker explores, he/she consumes near-accurate information about all of the patches but gives up the chance to exploit the most relevant patch for long. Understanding these operationalised scenario paves the way to model foraging behaviour in which user causes could be uncertainty, information overload, and confusion.

4 Quantum-inspired Reinforcement Learning Framework

We outline the proposed reinforcement learning approach to model the forager's action during an information seeking scenario where the task is to match a query for a given document in which the forager actions are queries. An agent interacts with its search environment characterised by a patchy distribution of information to find an optimal foraging strategy to maximise its reward. The forager's environment provides a fixed setting of optional information sources. Moreover, the forager has the choice to add a distinctive type of information patch into their diet. However, the distribution of distinctive information patches may consist

of information which the forager could likely not consume due to counterfactual situations in making decision amongst which patch (let us say document D1 and D2) contains certain information. In our framework, we consider the environment to be uncertain with dynamic parameters throughout a forager's search trail. The forager finds it difficult to differentiate patches and exploits experience to learn the environment. The increasing learning makes it complex at the dynamic and cognitive level where the forager's pursuit is to locate most relevant documents.

We use the *Actor-critic* policy gradient method [22] which inherently models such dynamics due to the forager's sequential behaviours that generate a continuous state representation. A forager's action (or state) can be described with the quantum superposition state and the corresponding updated state vectors, based on the respective interaction, can be achieved by random observation of the simulated quantum state based on the collapse principle of quantum measurement [28]. The probability of such an *action state vector* can be obtained by the probability amplitude which will be updated in parallel based on reward. This gives rise to new internal aspects in traditional RL algorithms which are policy, representation, action (in parallel) and operation update.

The quantum measurement decision process of a forager in selecting a document (the action) while seeking is ambiguous and uncertain [6]. In such situation, an observable describes possible actions (documents or information patches to select) and can be represented as ($\hat{O}$) with a base set containing $|0\rangle$ and $|1\rangle$ which corresponds to the two state vectors of $\hat{O}$. The measurement of a quantum system on the observable ($\hat{O}$) in a corresponding superposed quantum state ($|\psi\rangle$) refers to a measurement in superposition state. When making a measurement in state $|\psi\rangle$, the quantum state would collapse into one of its basis states $|0\rangle$ *or* $|1\rangle$. However, one cannot obtain a prior with certainty whether it will collapse to either of these states. The only information this quantum system can provide is $|0\rangle$ will be measured with probability $|\alpha|^2$ or $|\beta|^2$ as the probability to measure $|1\rangle$, where α and β represent the respective probability amplitudes.

We present a *quantum-inspired reinforcement learning (qRL)* framework for information seeking under dynamic search scenarios. The schematic architecture of qRL is shown in Fig. 1. qRL has two main components, an Actor-critic [22] based network to represent the RL agent which jointly encodes state and action spaces, and the information space known as environment containing documents. The Actor-critic components of an RL agent have their constructs subscribed via the Hilbert space formalism of Quantum theory [28].

Our framework is applicable to matching tasks, in particular, *semantic query matching* where candidate queries (extracted/predicted queries from the document) with the original document will be matched in a semantic Hilbert space (SHS) [33]. An SHS is a vector space of words, where words in combination involve a linear/non-linear formation of amplitudes and phases, delineating various level of semantics of combined words. In the SHS, a word w_i is represented by a base vector $|w_i\rangle$. Semantics of combined words are represented by superpositions

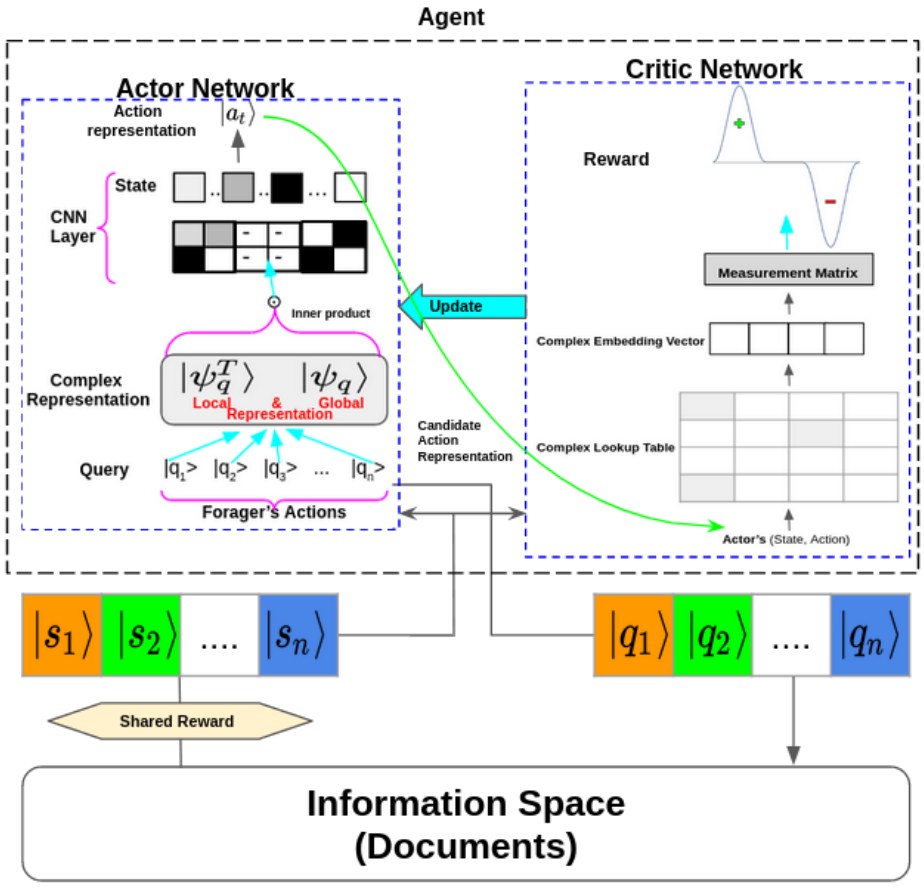

Fig. 1. Our proposed framework: Quantum-inspired Reinforcement learning-driven model for information seeker (in a semantic query matching task)

of word vectors, encoded in the probability amplitudes of the corresponding base vectors.

4.1 Preliminaries

The standard reinforcement learning is based on a finite state, discrete time Markov decision process (MDP) composed of five components: $s_t, a_t, p_{ij}(a), r_{i,a}$ and C, where s_t, the state at time t, delineates a_t the action at a specific time for a given state; $p_{ij}(a_t)$ is the probability of state transition (from state s_t to s_{t+1} via action a_t for all $t \in (i,j)$), r is a reward function where $r : \Gamma \longrightarrow \mathbb{R}$ with $\Gamma = \{(i,a) \mid i \in s_t,\ a \in a_t\}$, and C is an objective function.

In the following discussion we utilise tensor spaces. The notation in Table 1 follows those in [35, 34]. The fabric of our framework, i.e. the underlying Hilbert

space $\mathcal{H}$, is similar to the Tensor Space Language Model described in [34]. Here, the base vectors $\{|\phi_i\rangle\}_{i=1}^{n}$ of our n-dimensional space[4] are term vectors, either one-hot vectors or word embeddings. Any word vector $|w\rangle$ can be written as linear combination of the base vectors, i.e. $|w\rangle = \sum_{i=1}^{n} \alpha_i |\phi_i\rangle$ with $\alpha_i \in \mathbb{R}$ (or $\mathbb{C}$ in the complex case) as coefficient.

Table 1. Notations used in Reinforcement Learning Constructs, following [35, 34]. b_i depicts the dimension of orthonormal basis of Hilbert space, $\otimes$ depict the tensor product, R depicts the rank of $\mathcal{G}$ and $\mathcal{L}$ has n-order tensor of rank 1.

Notation	**Interpretation**	**Description**
α_{i,b_i}	$b_i \in \{1, ..., k\}$	Probability amplitude
$\|\phi_{b_i}\rangle$	Semantic meaning	Basis vector (n (word vectors) or k^n dimension for tensor product of basis vectors)
$\|w_i\rangle$	$\sum_{b_i=1}^{k} \alpha_{i,b_i} \|\phi_{b_i}\rangle$	Word state vector
$\|q_i\rangle$	$\|w_1\rangle \otimes \|w_2\rangle ... \otimes \|w_n\rangle$	Query state vector
$\|\psi_q^T\rangle$	$\sum_{b_1,...,b_n=1}^{k} (\underbrace{\prod_{i=1}^{n} \alpha_{i,b_i}}_{\mathcal{L}_{b_i,...,b_n}} \|\phi_{b_i}\rangle \otimes ... \otimes \|\phi_{b_n}\rangle)$	Local representation ($\mathcal{L}$ is a k^n dimensional tensor)
$\|\psi_q\rangle$	$\sum_{b_1,...,b_n=1}^{k} \mathcal{G}_{b_1 b_2 ... b_n} \|\phi_{b_i}\rangle \otimes ... \otimes \|\phi_{b_n}\rangle$	Global representation of combined meanings/patches
$\mathcal{G}$	$\sum_{r=1}^{R} w_r \cdot e_{r,1} \otimes e_{r,2} \otimes ... \otimes e_{r,n}$	Probability amplitude (semantic space of meaning)
$\langle \psi_q^T \| \psi_q \rangle$	$\sum_{b_1,...,b_n=1}^{k} \underbrace{\mathcal{G}_{b_1...b_n} \times \prod_{i=1}^{n} \alpha_{i,b_i}}_{Probability\ amplitudes}$	Projection of the global representation to the local representation of a query
State	$\prod_{i=1}^{n} \sum_{b_i=1}^{k} e_{r,i,b_i} \cdot \alpha_{i,b_i}$	Actor network state module (product pooling layer [35])
$\|a_t\rangle$	$(\|a_1\rangle, \|a_2\rangle, ..., \|a_R\rangle)^T$	Output of the Actor network

The overview of the RL process which possesses the Markov decision process is as follows:

Agent: In general, an agent acts as a controller within an information environment and is the one executing actions. In our framework, the RL agent is a forager (information seeker) available to the search environment (documents) delivering queries as actions, where the action is chosen based on the Actor-critic network (or the Policy network).

Action: An action a_t in the Web search scenario conforms to a query that searchers utilise to express their information need with the aim to either retrieve a document as an outcome of the query or continue the search process (exploratory search) (a

[4] The Hilbert space can be over the real or complex field, i.e. $\mathbb{R}^n$ or $\mathbb{C}^n$; we are assuming $\mathbb{R}$ for the further discussion

formal representation of the user action is shown in Table 1). In our framework, the forager (or searcher) action is to match a candidate query $|q\rangle$ (generated after inputting a set of queries) from document D to delineate $|q_{rD}\rangle$, where $|q_rD\rangle$ refers to a query state vector that represents the most optimal query for the selected document D given a positive/optimal reward (r). A candidate query is an outcome generated from the Actor network given the forager set of input queries.

State: A state s_t delineates the positive historical interaction of the forager with the search environment. In our framework, the Actor network has its *state* encoded by the product of the probability amplitudes of global-local projection $\langle \psi_q^T | \psi_q \rangle$ (of word meanings) for all words of a query We refer to the state representation defined by the product pooling method.

State Transition: The state representation describes the positive historical interaction of a forager. The transition among the states can be computed from the user's feedback. Our framework uses a convolutional neural network which has its convolution based on a state vector that encodes the historical interaction of the forager in finding the match of a query.

Policy: Policy is a strategic mechanism which represents the probability of a forager's action under a certain state. Our framework's policy network is stochastic, and we employ the Actor-Critic RL method [22] (Fig. 1) which assists the forager actions in the Actor network with a optimal policy value generated from the Critic. Thus, the Actor network estimates the probability of a forager action, and the Critic network gets the optimal value and updates it. The policy network is modelled as a probability distribution over actions and hence it is stochastic.

Reward: Reward ($r(s, a)$) in reinforcement learning is the success value of an agent's action a. This success value in information retrieval is interpreted in terms of the relevance judgement score [7]. Our framework process to receive reward values for the Critic network which inputs a pair of (state, action) that provides to the Actor network as an optimal reward for a given action which judges and scores the actions of the agent (or forager).

4.2 Our Proposed Framework

This fundamental RL definition is of utmost importance for proposing quantum-inspired reinforcement learning constructs. Following the quantum probability concepts, below are the constructs as follows (please also refer to Figure 1):

Actor Network: An Actor-critic [22] method refers to as policy gradient mechanism, where the Actor network for a given forager (or information seeker) in a particular state $|s_t\rangle$ outputs an action $|a_t\rangle$. This network inputs user queries (forager's actions), where these queries $|q_1\rangle, |q_2\rangle, ..., |q_n\rangle$ or a set of textual descriptions (which collectively form a document) form the local and global representations so as to model the interrelated interaction between words. Inspired by the notion of quantum theory, we employ the interpretation of wave function (due to the importance of word positions [36]) $|\psi\rangle$ as a state vector that can be explicated in RL constructs.

The Actor network inputs query state vectors $|q_1\rangle \ldots, |q_n\rangle$, where each word in a query is treated as a tensor product of vectors $|w_i\rangle$ and every word has a unique basis vector $|\phi_{b_i}\rangle$ that provides a generic semantic meaning with an associated probability amplitude. The speciality of a basis vector is that it can lead to a different meaning if interpreted severally across it. Then, we apply our framework in a semantic query matching task by a real-valued representation of queries by means of

local and global distribution so to allow such intermittent basis vectors that perceive the interaction between the meaning of different words. Hence, the wave function description of a query $|q_i\rangle$ can be depicted using the tensor product of words as $\left|\psi_q^T\right\rangle = |w_1\rangle \otimes |w_2\rangle ... \otimes |w_n\rangle$. A word dependency can be seen by expansion of tensors as $\left|\psi_q^T\right\rangle = \sum_{b_1,...,b_n=1}^{k} \mathcal{L}_{b_1...b_n} |\phi_{b_i}\rangle \otimes ... \otimes |\phi_{b_n}\rangle$, where $\mathcal{L}$ (the value is shown in Table 1) depicts the allied probability amplitude of k^n dimensional tensor in which it has the respective basis vectors $|\phi_{b_i}\rangle, ..., |\phi_{b_n}\rangle$ representing the meaning of the corresponding query. This tensor-based query representation is a local representation as a tensor with rank 1 actually delineates the local distribution of a query [34]. For words that are unseen in a query or compound meanings we need a global representation of them provided by a collective set of basis states (or vectors). A state vector (i.e., wave function of a query) to describe such a global representation is $|\psi_q\rangle = \sum_{b_1,...,b_n=1}^{k} \mathcal{G}_{b_1...b_n} |\phi_{b_i}\rangle \otimes ... \otimes |\phi_{b_n}\rangle$. This wave function delineates a semantic embedding space of n uncertain word meanings of a given query. The local and global representation differs in terms of their corresponding probability amplitudes i.e., $\mathcal{L}$ and $\mathcal{G}$, in which the probability amplitudes of the *global distribution* will be trained on a large collection of previous queries whereas the probability amplitudes of *local distribution* relates only to the input query. To compute the probability amplitudes among words from the input query (local representation) and unseen words generated from the global representation, we perform the inner product $\langle \psi_q^T | \psi_q \rangle$ of both representations that disentangle the interaction among it. The value of the projection is shown in Table 1. We use a convolutional neural network (CNN) to learn the obtained higher-dimensional tensor $\mathcal{G}$ (value shown in Table 1), where tensor rank decomposition can be used to decompose it (among other methods such as generalised singular value decomposition) and the decomposed unit vector $e_{r,n}$ with each rank 1 tensor of weight coefficient w_r. The unit vector is k-dimensional and the set of vectors $e_{r,n}$ acts as a subspace of tensor $\mathcal{G}$. The CNN inputs a query state vector with a convolution filter composed of the projection (inner product) among the $|q\rangle$ and the decomposed vector, which makes the CNN trainable. Then, the state representation (actor's state in Table 1) performs the product of all mapped unit vectors (from $\mathcal{G}$) for all the subwords of a query. After all these operations, the Actor network yields an action state vector $|a_t\rangle$ (action a_t at time t) to depict a set of matched words.

Critic Network: The Critic network of the qRL framework is based on a quantum-like language model parameterised CNN which inputs the generated state and the candidate action $|a_t\rangle$ from the Actor network. The output of the Critic network is a scalar value or value of the Q-function [10]. The reward values $R_e \in [-1, 0, 1]$ reflect the ability of the candidate action generated by the Actor network. The significance of the reward value represents the probability of designating the correct label to action i.e, the multi-class classification of queries to match among documents will be used to update the reward. Rewards (or classification labels) are categorised as -1 for a mismatched query which has negative word polarity (leading to a compound meaning). For instance, "dogs chase cats" and "dogs do not chase cats" contribute to a compound meaning itself but in an opposite sense. We tend to consider that a word renders the entire polarity of a query, provided to which new word it associates with. A realistic example of this hypothesis can relate to one of our framework's main constructs i.e., $|q\rangle$ which is a state vector equal to the tensor product of possible words, where the word coefficients (i.e., probability amplitudes) of basis vectors can be altered to derive a new query giving rise to a compound meaning. The negative word polarity example is an

actual scenario of it. Positive and zero rewards are classed as matching and partially matching for queries.

In the Critic part, the concatenation of the actor's state and candidate action is performed using one-hot encoding in which the query is passed via a complex-valued lookup table, where each word in their own superposition state is encoded into a complex embedding vector [32]. Then, a measurement is performed using the square projection to compute the query density matrix from the complex embedding vectors. The probability of a measurement can be estimated using the Born's postulate for a given query state ρ (a density matrix) which is $\mathbf{p} = \mathrm{Tr}(P\rho)$, where $\mathbf{p}$, P and Tr represents the class of the query, projection matrix, and trace of a matrix, respectively. The density state $\rho = \sum_{i=1}^{n} \beta_i \left|w_i\right\rangle \left\langle w_i\right|$ of a query is perceived as the word states in combination, provided that the density matrix ($\left|w_i\right\rangle \left\langle w_i\right|$) reflects a word ($w_i$) in superposition state (in this case $\sum_{i=1}^{n} \beta_i = 1$). The generated query density matrix has its diagonal and non-diagonal entries as real-valued and complex nonzero values, and both type of entries inherently inform about the distribution of semantic and contingent meanings. We adopt the interpretation of complex phase introduced in [32] to compute the sentence density matrix which has word senses as positive, neutral, and negative. The reward is estimated using such interpretation from the measurement matrix. A pictorial representation of the Critic network is shown in Fig. 1.

In brief, the Actor-Critic policy network helps suffice the number of components with respect to traditional reinforcement learning. Also, the Agent part of the framework acts as a controller for the user in the same way Information Foraging mechanisms possess to a searcher. IFT helps a searcher through suggesting an optimal foraging path via information scent, and here in the framework the Critic network informs/updates the Actor with a value (reward) for a certain action that is positively rewarded. Hence, our framework meets foraging in certain regards (such as information seeking behaviour assessed as foraging and inherently as RL task).

Rewards: The forager aim is to identify the relevant match (or a perfect match) of a query (or patch) for the clicked/selected document that can be perceived as its reward. However, our framework's reward function is designed in a way to guide the forager on how to perceive the document information and draw the most relevant match (patch). Also, the reward value is discrete as it revolves around -1, 0, and +1. The definition of reward in reinforcement learning [10][5] resembles a certain analogy of information scent, which is a measure of utility and results in two types of information scent score – a scalar value and the probability distribution of scent patterns [18]. In RL, the perspective of value distribution of received reward by an agent can depict the analogous nature of information scent patterns. Hence, an explainable approach of reinforcement learning-based rewards using the IFT-based model of information scent can give further intuition to negative rewards. Information scent can be interpreted as the perceived relevance of rewarded actions defined as positive and negative scent values. The physical meaning of positive and negative information scent scores are that the forager accumulates rich information along the path he/she foraged to locate the relevant information, and the unhealthy consumption of information reckon searcher negative towards the search environment which leads them to give up the information world (or RL environment) or task itself.

Update Probability Amplitude/Policy: To update the probability amplitude in the Actor network, the important part is to measure the actions for some certain

[5] Rewards can be normalised to generate outcomes in reinforcement learning [10]

states which on collapse will give rise to the occurrence probability of the norm of state vector for the particular candidate action, which later will execute the Actor network. The more we record the experience and learning of each action (even erroneous action), the probability amplitude becomes more informative. We know that the action $|a_t\rangle$ is the tensor product of all possible words and to calculate one user action (i.e. $|a\rangle$) from it can be possible while interacting with changes in probability amplitudes for the combined meaning.

5 Conclusion and Future Work

In this paper, we propose a mathematical framework of reinforcement learning inspired by the Hilbert space formalism in Quantum theory. The framework models the learning process of forager actions in a semantic query matching task given the search environment is patchy. The core of our framework is to characterise a forager with very little or unclear information about their search pattern, unclear or evolving information need and features. Also, no information about how a forager makes their trail (initially the information scent is unknown and emanates as it follows via distinct cues) choice during finding information and the amount of information they consume in real-time interaction with the search system. Apart from this, the major trade-off situation of exploration and exploitation in the foraging process makes the process of understanding about the forager's search actions complex. To tackle such a complex process of dynamic action for a state and vice-versa, we adapt the Actor-critic reinforcement learning method as a policy network, in which the actor network is continuously informed about the generated action from the critic network. The framework subscribes to the quantum probability constructs to model by the representation of forager search actions and states. Quantum theory has been earlier applied in the area of information seeking [6], but representing and measuring actions of each state is a challenging scenario due to the continuous update of state-action in parallel, so using the Actor-critic reinforcement learning method paves the way to influence learning and representation mechanisms; many complex IR problems could be interpreted appropriately in a new way within such an inclusive framework. In the future, we intend to evaluate this framework for certain IR tasks.

Acknowledgements

This work is part of the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 721321.

References

1. Pirolli, P., & Card, S. (1995, May). Information foraging in information access environments. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 51-58).
2. Chowdhury, S., Gibb, F., & Landoni, M. (2011). Uncertainty in information seeking and retrieval: A study in an academic environment. Information Processing & Management, 47(2), 157-175.

3. Tran, V. T., & Fuhr, N. (2012, August). Using eye-tracking with dynamic areas of interest for analyzing interactive information retrieval. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 1165-1166).
4. Pirolli, P., & Card, S. (1999). Information foraging. Psychological review, 106(4), 643.
5. Brennan, K., Kelly, D., & Arguello, J. (2014, August). The effect of cognitive abilities on information search for tasks of varying levels of complexity. In Proceedings of the 5th Information Interaction in Context Symposium (pp. 165-174). ACM.
6. Wittek, P., Liu, Y. H., Darányi, S., Gedeon, T., & Lim, I. S. (2016). Risk and ambiguity in information seeking: Eye gaze patterns reveal contextual behavior in dealing with uncertainty. Frontiers in psychology, 7, 1790.
7. Tang, Z., & Yang, G. H. (2019). Dynamic Search–Optimizing the Game of Information Seeking. arXiv preprint arXiv:1909.12425.
8. Piwowarski, B., Frommholz, I., Lalmas, M., & Van Rijsbergen, K. (2010, October). What can quantum theory bring to information retrieval. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 59-68).
9. Charnov, E. L. (1976). Optimal foraging, the marginal value theorem.
10. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
11. Chandak, Y., Theocharous, G., Kostas, J., Jordan, S., & Thomas, P. (2019, May). Learning Action Representations for Reinforcement Learning. In International Conference on Machine Learning (pp. 941-950).
12. White, R. W. (2014). Belief dynamics in Web search. Journal of the Association for Information Science and Technology, 65(11), 2165-2178.
13. Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001, March). Using information scent to model user information needs and actions and the Web. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 490-497). ACM.
14. Chi, E. H., Pirolli, P., & Pitkow, J. (2000, April). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 161-168). ACM.
15. Liu, H., Mulholland, P., Song, D., Uren, V., & Rüger, S. (2010, August). Applying information foraging theory to understand user interaction with content-based image retrieval. In Proceedings of the third symposium on Information interaction in context (pp. 135-144). ACM.
16. Jaiswal, A. K., Liu, H., & Frommholz, I. (2019). Effects of Foraging in Personalized Content-based Image Recommendation. arXiv preprint arXiv:1907.00483.
17. Jaiswal, A. K., Liu, H., & Frommholz, I. (2019, December). Information Foraging for Enhancing Implicit Feedback in Content-based Image Recommendation. In Proceedings of the 11th Forum for Information Retrieval Evaluation (pp. 65-69).
18. Jaiswal, A. K., Liu, H., & Frommholz, I. (2020, April). Utilising information foraging theory for user interaction with image query auto-completion. In European Conference on Information Retrieval (pp. 666-680). Springer, Cham.
19. Nogueira, R., Bulian, J., & Ciaramita, M. (2018). Learning to coordinate multiple reinforcement learning agents for diverse query reformulation. arXiv preprint arXiv:1809.10658.

20. Du, J. T., & Spink, A. (2011). Toward a web search model: Integrating multitasking, cognitive coordination, and cognitive shifts. Journal of the American Society for Information Science and Technology, 62(8), 1446-1472.
21. White, R. W., Bennett, P. N., & Dumais, S. T. (2010, October). Predicting short-term interests using activity-based search context. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1009-1018). ACM.
22. Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Abbeel, O. P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In Advances in neural information processing systems (pp. 6379-6390).
23. Zhang, B. T., & Seo, Y. W. (2001). Personalized web-document filtering using reinforcement learning. Applied Artificial Intelligence, 15(7), 665-685.
24. Seo, Y. W., & Zhang, B. T. (2000, January). A reinforcement learning agent for personalized information filtering. In Proceedings of the 5th international conference on Intelligent user interfaces (pp. 248-251). ACM.
25. Eliassen, S., Jørgensen, C., Mangel, M., & Giske, J. (2007). Exploration or exploitation: life expectancy changes the value of learning in foraging strategies. Oikos, 116(3), 513-523.
26. Yue, Y., & Joachims, T. (2009, June). Interactively optimizing information retrieval systems as a dueling bandits problem. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 1201-1208). ACM.
27. Balabanović, M. (1998). Exploring versus exploiting when learning user models for text recommendation. User Modeling and User-Adapted Interaction, 8(1-2), 71-102.
28. Van Rijsbergen, C. J. (2004). The geometry of information retrieval. Cambridge University Press.
29. Von Neumann, J. (2018). Mathematical Foundations of Quantum Mechanics: New Edition. Princeton university press.
30. Fakhari, P., Rajagopal, K., Balakrishnan, S. N., & Busemeyer, J. R. (2013). Quantum inspired reinforcement learning in changing environment. New Mathematics and Natural Computation, 9(03), 273-294.
31. Zhou, J., & Agichtein, E. (2020, April). RLIRank: Learning to Rank with Reinforcement Learning for Dynamic Search. In Proceedings of The Web Conference 2020 (pp. 2842-2848).
32. Li, Q., Uprety, S., Wang, B., & Song, D. (2018). Quantum-inspired complex word embedding. arXiv preprint arXiv:1805.11351.
33. Wang, B., Li, Q., Melucci, M., & Song, D. (2019, May). Semantic Hilbert space for text representation learning. In The World Wide Web Conference (pp. 3293-3299).
34. Zhang, L., Zhang, P., Ma, X., Gu, S., Su, Z., & Song, D. (2019, July). A generalized language model in tensor space. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 7450-7458).
35. Cohen, N., Sharir, O., & Shashua, A. (2016, June). On the expressive power of deep learning: A tensor analysis. In Conference on Learning Theory (pp. 698-728).
36. Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., & Simonsen, J. G. (2019, September). Encoding word order in complex embeddings. In International Conference on Learning Representations.