Case-Based Sample Generation using Multi-Armed Bandits

Andreas Korger[1,2] and Joachim Baumeister[1,3]

[1] University of Würzburg, Am Hubland, Würzburg, D-97074, Germany
[2] a.korger@informatik.uni-wuerzburg.de
[3] joba@uni-wuerzburg.de

Abstract. A central problem in knowledge-based tasks is to provide a collection of reusable knowledge samples extracted from a textual corpus. Often, such corpora are structured into different documents or topics, respectively. The samples need to be proven for usability and adapted by a domain expert requiring a certain processing time for each sample taken. The goal is to achieve an optimal retrieval and adaptation success meeting the time budget of the domain expert. In this work, we formulate this task as a constrained multi-armed bandit model. We combine it with the model of a configurable data-driven case-based learning agent. A case study evaluates the theoretical considerations in a scenario of regulatory knowledge acquisition. Therefore, a data set is constructed out of a corpus of nuclear safety documents. We use the model to optimize the evaluation process of sample generation of adaptational knowledge. The corresponding knowledge graph has been created in an information extraction step by automatically identifying semantic concepts and their relations.

Keywords: Case-Based Reasoning · Multi-Armed Bandits · Agent-Based Modeling · Semantics · Knowledge Management · Sampling.

1 Introduction

Let us consider the following situation: a domain expert needs to write a new safety document. He has a corpus available with a collection of documents similar to a safety domain. So he may reuse knowledge contained in the corpus and collect new safety knowledge by searching the existent corpus. He adapts promising textual passages to his needs and drops others. In complex domains an unknown document is like a black box which has to be understood, even for domain experts. Therefore, he has to analyze and interpret passages of the documents in the corpus, assess their quality, and adapt them to his specific domain. In the end, if the expert finds the knowledge he is looking for, he is satisfied with the selection of documents he made. This simple process is complicated by various factors.

The expert has special characteristics. He has some prior knowledge and he has limited time. Subsequently, he expects to find the knowledge in the corpus

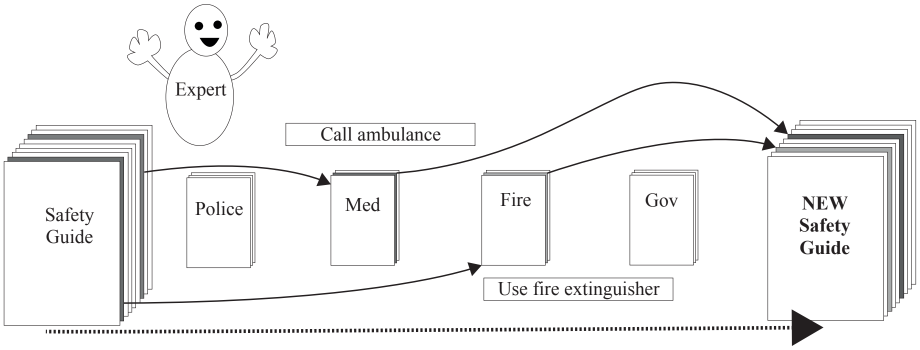

Fig. 1: Retrieval and adaptation of existent textual samples to support an expert in the creation of a new safety document.

he is searching for, within the time budget he has available. Additionally to the time budget the expert is characterized by a certain preference. He is interested in certain topics. Despite that, the expert hopes to find new helpful knowledge he does not yet know. A fact, that leverages his task, is that the corpus is structured by domain experts into documents representing a self-contained knowledge scheme as illustrated in Figure 1. In summary, every document of the corpus has a certain quality which is initially unknown to the reader and comes up step by step with every textual sample processed. The following exemplary textual sample is taken from a document for nuclear safety which is part of the evaluating case study presented in Section 4.

(1) **Example:** The *staff assigned* the *responsibility* for carrying out such *reviews* for issues of *fire safety* should be suitably *qualified* to *evaluate* the potential effect of any *modification* on *fire safety* and should have sufficient *authority* to *prevent* or *suspend* modification work, if necessary, until any *issues* identified have been satisfactorily *resolved*.

The task of retrieving good new adaptation candidates inherits two competing goals. On the one hand, the expert needs to distribute his search in the corpus to find the documents that fit his expectations (*explore corpus*). On the other hand, when he found a good document, he does not want to waste time (*exploit good document*) searching for other documents, that may fit his expectations better. A well accepted strategy to model the before described scenario of sequential resource (*time*) distribution amongst competing alternatives (*documents*) is the bandit model [21]. The case-based paradigm, that similar problems (*retrieved textual passage*) have similar solutions (*adapted textual passage*), aids to connect the bandit model with the characterization of the expert [7]. The expert is modeled as a data-based agent [22]. The character of the agent regarding his prior knowledge, preferences, and learning goal are modeled as a case base together with a configurable similarity model.

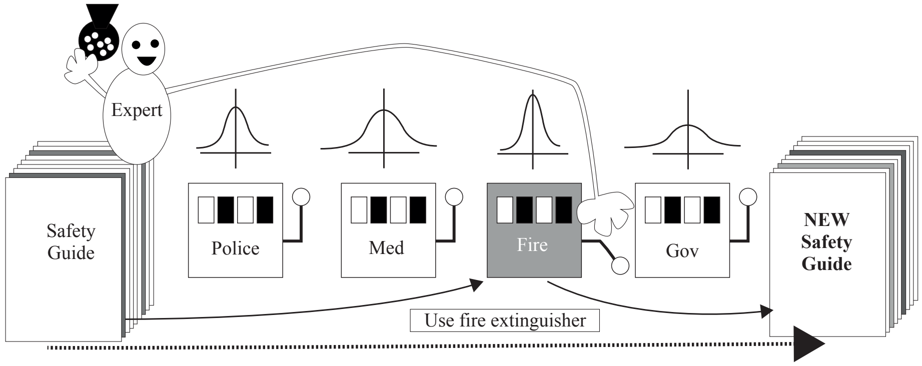

Fig. 2: Representation of the textual sampling process as a bandit model. Every document is represented as one bandit arm having a certain mean and variance of quality distribution for the contained information units (cases).

1.1 Problem Description

The described scenario is connected with the solving or mitigating of the following problems. The *exploration-exploitation dilemma* of using already visited "good" documents for sample generation rather than sampling completely unknown or allegedly "bad" documents. How can a sequential reward model be constructed that bases solely on the existent data. Can the model be used to calculate the quality of a retrieved sample depending on the characteristics of the agent and thus defining "good" and "bad" quality of documents.

1.2 Solution Approach

We facilitate and formulate the scenario as a constrained multi-armed bandit model which offers strategies to mitigate the *exploration-exploitation dilemma* in a configurable way. We use case-based reasoning (CBR) strategies incorporating an initial agent setup (*initial case base*) together with a learning goal (*optimum case solutions*) to construct a configurable reward model on the base of *similarity assessment*. We use statistical language models to calculate similarity components of the retrieved information for *sample adaptation*.

1.3 Contribution and Research Question

With the presented approach we aim to answer the following research question. Is it possible to formulate the sampling process of a textual corpus as a multi-armed bandit problem in combination with a data-driven agent characterization. Can the phrases (*samples*), documents (*bandit arms*), and corpus (*agent environment*) thereby be considered as discrete semantic unities even though they are interrelated. Do the documents have a mean quality and variance of quality regarding the sampling process of contained phrases as depicted in Figure 2.

2 Related Work

There exist several works describing document retrieval or ranking using the strategy of multi-armed bandits. Perotto et al. [18] use bandits for document retrieval in the juridical domain. They incorporate the searching characteristics coded in queries done by previous users to leverage the performance for the current query. While this work has interesting ideas to use past user behavior, it differs in its structure in that way that it bases on single queries and lacks the integration of an agent based user characterization. Losada et al. [17] propose a bandit-based pooling strategy for document adjudication. A combination of active learning and multi-armed bandits is proposed by Rahman et al. [20] with the intent of selecting the best document for a testing collection for evaluation purposes. Most approaches focus on document selection or ranking wheres we focus on sample selection and just exploit the documental clustering.

Related work that supports the construction of a convenient agent model treated the following topics. A resource oriented variation of the multi-armed bandit problem is presented by Bengs and Hüllermeier [6]. It aims to minimize the resource limit and the risk of exceeding resources. The idea of introducing a learning goal is also picked up by Brändle et al. [8]. Racharak et al. [19] present a a concept-based similarity measure that incorporates the preferences of an agent in description logic. Insight into the relation between human psychology and the multi-armed bandid strategy was given by Schulz et al. [23]. Their work investigates how a latent structure in the bandit task is connected to the natural learning behavior of a searching agent.

Concerning variants of multi-armed bandit modeling an outstanding approach is the hierarchical structuring of the action space of the agent. As documents are most often structured into a hierarchy of topics this can yield significant improvements in the overall performance. Especially, as the here presented case study relies on hierarchical structuring of semantic concepts synergies should not be neglected. Hong et al. [12] present fundamental considerations in this direction. Kumar et al. [15] use a hierarchical bandit model in combination with decision tree algorithms for the identification of users in social networks having special attributes. This problem setting shows similar characteristics as the hierarchical structuring of attributes is similar to the hierarchical structuring of semantic concepts. Carlsson et al. [9] show how a clustered structure of bandit arms can be exploited to improve the Thompson sampling strategy. Important aspects of linked data in connection to non stochastical bandit modeling are addressed in the work of Alon et al. [2]. Basic considerations about non-stochastical bandit models have been presented by Auer et al. [4]. A different hierarchical modeling approach is presented by Sen et al. [24]. They chose to represent the problem with hierarchically structured arms.

We will address, integrate, and extend different aspects of these related approaches into a new combination of agent-bandit model which is explained in the following.

3 Setting

In the following we will formalize the introduced problem description. We start with assumptions that are made to facilitate the scenario. We present and explain the ideas behind formal definitions of the learning agent and the reward model. We outline solution strategies using the presented setting.

3.1 Assumptions

The textual corpus is analyzed in a natural language processing step. The documents are chunked into *uniform informational pieces* of a certain meaningful size. Such units can be retrieved as samples, for instance, a paragraph or sentence. The textual corpus is enriched by *semantic meta knowledge* identifying and annotating safety related semantic concepts and their relations. We assume a closed world, namely, the textual corpus together with its semantic representation. This means that the learning agent is provided with some *prior knowledge* that is part of the textual corpus and its semantic representation. We assume a given *learning goal* which is also part of the textual corpus but is hidden to the agent. The *preferences* of the agent are provided by a set of semantic concepts defined on base of the semantically annotated textual corpus saved in a knowledge graph $\mathcal{O}$. It is unknown what would be an ideal learning goal. Therefore, we assume a subset of the corpus as learning goal. The agent is informed about the fulfillment of the goal by similarity information given to him by a hypothetic teacher. There are some flaws in this modeling: The prior knowledge increases with every sample processed by the agent and the learning goal decreases if partially met. This would lead to a non stationary reward model which changes over time. For simplification we assume a stationary model with the same reward configuration over the whole sampling process.

3.2 Formal Representation

We consider a corpus $\mathcal{C} = \{D_1, ..., D_m\}$ divided into m documents each consisting of $I_n, n \in \{1, .., m\}$ information units with $i_i \cap i_j = \emptyset \; \forall i \neq j$. Let $\mathcal{I}_\mathcal{C} = \{I_1, .., I_m\}$ be the set of all information units contained in the corpus. This corpus is represented as a K-armed bandit $\mathcal{K} = \{1, ..., K\}, K = |\mathcal{C}|$ with a set of K arms where each arm stands for one document. The agent is willing to read a number of b retrieved passages of text. By the action a_i of pulling the arm k at time t a sample piece of the document k is provided which is denoted as A_t. The quality of the sample piece generated by this action is the reward $R_k \in \{0, 1\}$ (with $R_k = 0$ meaning sample rejected and $R_k = 1$ meaning sample accepted). With ongoing time steps a sequence of actions $(A_1, A_2, .., A_b)$ with according rewards $(R_1, R_2, .., R_b)$ is produced. The (discrete) time goes on until the budget of the expert is consumed and he is not willing to take more sampling actions. What we are searching for is the optimal combination of actions to make the best out of the experts budget. The ”expert” is formalized as an agent with the following characteristics.

Definition 1 (Learning Agent Scenario). *Let $\mathcal{E} = \langle \mathcal{P}, \mathcal{G}, \mathcal{L}, b \rangle$ be an agent that is modeling an expert with a prior knowledge $\mathcal{P} \subset \mathcal{I}_{\mathcal{C}}$, with the preferences $\mathcal{L} \subset \mathcal{O}$ regarding topics he is interested in. We define a learning goal aligned to the agent as a subset $\mathcal{G} \subset \mathcal{I}_{\mathcal{C}}$. The agent has a budget of b samples that he is willing to take, meaning the bandit game goes T rounds ($T = b$). The environment of the agent is the textual corpus. The agent has the actions of pulling a bandit arm, accepting a sample, and rejecting a sample.*

Altogether, the information units that define the agent are considered as the initial case base. Each sample is considered as a new case, potentially representing a (partial) solution for the agent's task. If samples are indeed adaptable, is decided by comparing the new solutions to the existing case base. Finally, the agent is capable of solving his problem of creating the new document.

(2) **Example:** An exemplary corpus consists of four different safety documents. With the topics "fire safety", "police guidelines", "medical guidelines", and "governmental security" as depicted in the Figures 1 and 2. Each document consists of 1,000 relevant phrases and the expert has a budget of 100 samples, which he is capable of reading and analyzing. His preferences are "transportation" and "mobile fire safety" and he wants to create a document for the safety of utility vehicles in factories. He is a governmental fire safety expert and his prior knowledge is a subset of 50 phrases each of the fire safety and the governmental document. How does he distribute his 100 samples over the 4,000 existing phrases providing the best sampling success to him.

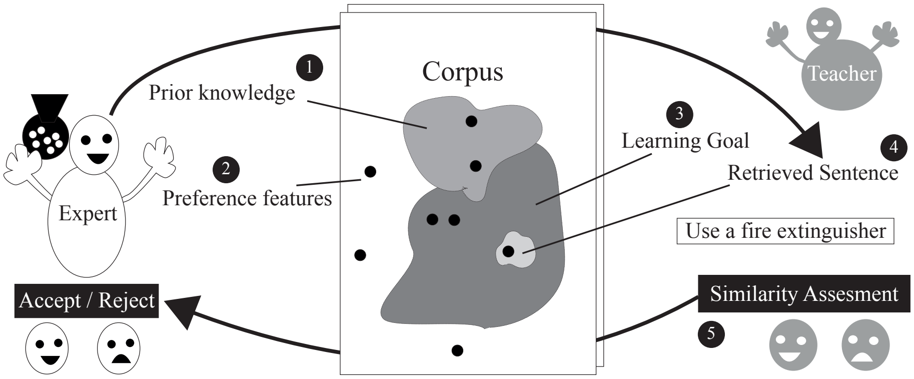

Fig. 3: Case-based cycle of corpus sample retrieval with joined agent-teacher similarity assessment. The expert has some prior knowledge (1) and preference features (2). The teacher defines a learning goal (3) which is hidden to the expert but gives him feedback about the quality of retrieved samples (4). Via a similarity assessment (5) the expert accepts or rejects the sample, then adapts it to his needs and retains it in the new corpus as a new case.

3.3 Reward Model

For every sampling action the agent is rewarded. To create a model for this reward we exploit the case-based paradigm that similar problems have similar solutions [7]. We therefore construct a similarity model that calculates the similarity between the retrieved sample (new case) and the characteristics of the agent defined by the agent model (case base). The higher the similarity the better the sample quality. If this numerical quality lies above a certain threshold the agent accepts the sample for adaptation, otherwise it is rejected. An illustration of this process is depicted in Figure 3.

Definition 2 (Reward Function). *Let sim(A_t,$\mathcal{G}$) be the similarity of the retrieved sample as action A_t with the learning goal, sim(A_t,$\mathcal{L}$) the similarity between the retrieved sample and the preferences of the agent, and sim(A_t,$\mathcal{P}$) the similarity between the sample and the prior knowledge. Then the reward function $R(A_t) \in \{0, 1\}$ with $\alpha, \beta, \gamma \geq 0$ is defined as:*

$$sim_R(A_t) = \frac{\alpha sim_X(A_t, \mathcal{G}) + \beta sim_X(A_t, \mathcal{L}) + \gamma(1 - sim_X(A_t, \mathcal{P}))}{\alpha + \beta + \gamma} \tag{1}$$

$$R(A_t) = \begin{cases} 1 & for\ sim_R(A_t) \geq \delta \\ 0 & for\ sim_R(A_t) < \delta \end{cases} \tag{2}$$

$$sim_X(A_t, X) = \max(Sim_{SLMF}(A_t, I) \forall I \in X) \tag{3}$$

$$sim_{SLMF}(A_t, I) = sim_{SLM}(\frac{\sum_{k=1}^{m} f_j}{m} \forall f \in A_t, \frac{\sum_{l=1}^{n} g_l}{n} \forall g \in I) \tag{4}$$

We define a stacked reward function basing on the similarity assessment weighting and averaging the similarities to preferences, learning goal, and the dissimilarity to the prior knowledge. The reward model is constructed in a heuristic way. The local similarities $sim_X(A_t, \mathcal{G})$ and $sim_X(A_t, \mathcal{L})$ add to each other because both a similarity to the learning goal and the preferences of the agent are desirable. If there is a similarity to the prior knowledge of the agent this is considered as not desirable because the agent prefers to learn new knowledge. If $sim_X(A_t, \mathcal{P})$ is zero then R is not influenced, a high similarity of A_t and P leads to a reward of 0 adjustable by the weights γ and δ. The similarities between A_t and $\mathcal{G}, \mathcal{L}, \mathcal{P}$ are calculated using a statistical language model with feature focus (SLMF) [27]. The selected sample is pairwise compared to every information unit of $X \in \mathcal{G}, \mathcal{L}, \mathcal{P}$. The highest similarity value is taken as the value of $sim_X(A_t, X)$. The feature-based similarity $sim_{SLMF}(A_t, I)$ for all selected features $f, g \in A_t, I$ is calculated by taking the average of the embedded feature vectors [27]. Which features are selected is left as a hyper-parameter of the sampling process. Selected features could be for instance only nouns, relevant safety features, only verbs, specific relations, etc [13].

3.4 Solution Strategies

Solution strategies for bandit models use different approaches to find a balance of exploiting good bandit arms already visited and exploring unknown bandit arms. We present and evaluate three approaches to the presented setting. The naive strategy of selecting randomly will be used as a baseline for the evaluation of the heuristic strategies. Let $\mu_k \in [0,1]$ be the mean reward of pulling an arm k at time step t-1. This is the expected reward for the next time step t for this arm which is denoted as $Q_t(a_k) = \mathbb{E}[R_t \mid A_t = a_k]$. In a "real" stochastic setting, if an arm was pulled infinite times the true mean value would be known. If $Q(a_k)$ is calculated for the discrete number of all information units contained in a document then the expected mean reward of the document is known. If the agent would know this hidden quality of all documents he could choose an optimal sequence of actions $a_* = \text{argmax}_a \mathbb{E}[R_t \mid A_t = a_k]$. Which will later be used to evaluate the performance of solution strategies.

Epsilon-Greedy Strategy A baseline strategy to address the exploration-exploitation dilemma is the epsilon-greedy strategy (EG) [25]. The parameter epsilon (ε) defines how eager the agent is for exploration. The higher the value of ε the more unknown documents will be visited. At each time step an unknown document out of $\mathcal{D}$ is chosen with a probability of ε for sample retrieval. At a probability of $1 - \varepsilon$ the document is selected for the next retrieval out of which the sample with the maximum reward so far was generated: $A_t = \text{argmax}_a Q_t(a)$ which is the "greedy" or exploiting component of the algorithm.

Upper Confidence Bound Strategy The epsilon-greedy strategy takes samples from random documents at a constant percentage. This neglects that in the later time steps there is already knowledge about the environment available. The Upper Confidence Bound strategy (UCB) makes use of this knowledge and changes the ratio of exploration and exploitation [3]. This is achieved with a bias added to the actually expected mean value of an action which decreases with increasing number of pulls of the according bandit arm. The greedy step changes to: $A_t = \text{argmax}_a [Q_t(a) + c\sqrt{\frac{\log t}{N-t(a)}}]$ where $N_t(a)$ denotes the number a document has been already selected for sampling and c is a parameter which controls the ratio of exploration, the bigger c the more exploration is done.

Thompson Sampling Strategy The Thompson sampling strategy (TS) [26] differs from the previous approaches. From the received rewards a probability model is calculated for each bandit arm and refined with every sample received. These probability models are then used to decide which action to take best. In the present setting actions are considered to have only two outcomes, sample accepted or sample not accepted. This binary reward scenario can be described by a Beta distribution which approximates the behavior of each bandit arm [26]. The so far made theoretical considerations are used in practical application in the following case study.

4 Case Study

The spark for this work developed out of the task of evaluating the quality of an automatically populated ontology. The available time budget allowed for a maximum analyzation capacity of some hundred annotated textual samples. Compared to a dataset size of more than 200,000 samples a better strategy than random sampling was necessary. Additionally, a configurable sampling setup was desired depending on the task and user profile.

For the present experimental evaluation we use a dataset created from a textual corpus published in the domain of nuclear safety. This corpus was previously annotated and transformed into an according dataset of about 222,000 sentences. The corpus consists of publicly available 143 documents containing in total about 14,000 pages of English text. These documents were published by the IAEA (International Atomic Energy Association), which is a sub organization of the United Nations [1]. The IAEA aims to regulate the domain of nuclear safety on an international level and gives advise and support to national authorities. Table 1 gives an insight into selected subjects the documents of the corpus are aiming to regulate.

Table 1. Selected documents of the corpus.

Number	Document Title
(1091)	“Fire Safety in the Operation of Nuclear Power Plants”
(1159)	“External Events Excluding Earthquakes in Nuclear Power Plants”
(1191)	“Protection against Internal Hazards other than Fires”
(1798)	“Regulations for the Safe Transport of Radioactive Material ...”
(1368)	“Predisposal Management of Radioactive Waste”
(1546)	“Nuclear Security Systems and Measures for Major Public Events”

4.1 Semantic Fundamentals

Additionally to the plain documents, a terminology is published and maintained by the IAEA covering about 1,500 semantic concepts of the domain using the RDF data model [29] and the SKOS standard for knowledge organization [28]. The terminology is structured into several hierarchical layers and contains concepts like `:fire`, `:manualFireFighting`, and `:fireProtection` together with definitions and explanations. Out of these concepts, *incidents* and *safety measures* where identified using lexico-syntactic patterns and an open information extraction approach [11, 5]. The retrieved information was annotated using the RDF-star data model [10] and saved in a knowledge graph [13, 14].

Example 3 shows a phrase extracted from the Document 1191 listed in Table 1. Because of the concepts ”vessel” and ”fuel” this phrase is similar to the ”utility vehicle” scenario described in Example 2. This example points out the time-consuming and pseudo-stochastic nature of the adaptational process. The success of having retrieved a good passage is not guaranteed. For instance, ”vessel” is here used in the sense of a ”container” and not a vehicle-like object. Nevertheless, concepts seem similar and the phrase might be promising. To proof and adapt this phrase the expert needs to research, what should be done in this context for the incident of a ”missile”, what is meant by ”special design features”, then proof whether this would be a good strategy for his own scenario, maybe consult other experts, and finally rewrite the phrase.

(3) **Example:** For reactors equipped with vessel closure plugs to retain the fuel in position, special design features should be provided to ensure that the probability of *ejection of the closure plug* is low. In the absence of such special features, the consequences of the *failure or the ejection of a single closure plug* should be evaluated as for a *missile*.

Beyond that, the semantic knowledge can be used for the purpose of document filtering. It is not necessary to use all information units in a document. In a filter step only those units relevant for a certain task can be pre-selected. For instance, only sentences that contain a certain type of relation, that fall under a certain topic, and show other distinct characteristics. A second benefit is the availability of meta-knowledge about features to calculate similarity functions and adapt retrieved information units to different scenarios [7]. Furthermore, it is not obligatory to stay in the document-based clustering of information units. The bandits can also be setup using other arbitrary clustering approaches. For instance, by creating one bandit arm for each class of relations available, for a selection of topics, specific case attribute oriented clusters, and algorithmic provenance of annotation. In this manner the approach generalizes to a variety of possible application scenarios.

4.2 Experiments

In the following we present a collection of experiments that describe determination of hyper-parameters, individual characteristics of each solution policy, and comparable aspects of the bandit problem solution strategies. We compare the experiments against a baseline of random sampling and use the concept of *regret* for performance evaluation. The dataset was divided into an experimental data environment with training, validation, and test splits, leaving 10% of all sentences for each; validation and testing. We initially tested the algorithms with a small selection of 3 documents, that where known in terms of document content to investigate the behavior with human insight. We then scaled the algorithms to a number of 10 and 100 documents. We used two different agent configurations A_1 and A_2 as described in Figure 4.

Initialization, Hyper-Parameters and Optimal Strategy For the EG and the UCB algorithms we initialize all arms with a mean reward of 1. This ensures that each arm is visited at least once. For the same reason for the TS policy

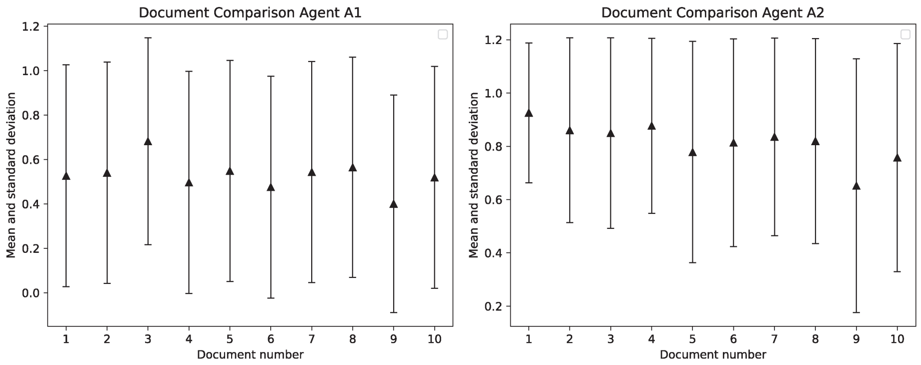

Fig. 4: The figure shows the mean of the reward for each document together with the standard deviation for a fixed reward configuration for ten documents of the corpus amongst which the documents mentioned in Table 1. With $\alpha = 1$, β=0, γ=0, δ=0.8 for Configuration A_1 where just the learning goal was defined as a fixed single sentence with $\mathcal{G} = 1$, $\mathcal{P} = 0$ sentences and $\mathcal{L} = []$. For Configuration A_2 and $\alpha = 1$, β=1, γ=1, δ=0.5, a fixed set of $\mathcal{G} = 17$ random sentences sampled from three documents, $\mathcal{P} = 1$ fixed sentence , and $\mathcal{L}$ the preferred semantic concepts `:fire`, `:transportation`, and `:leakage`.

the distribution of each arm is initialized with a count of positives = 1 and negatives = 1. These starting values determine a wide spread initial distribution. To determine a reasonable range of time steps a human expert would accept, we considered the following. We estimated about five minutes of time to manually execute the adaptational steps needed in case of Example 3. The documents

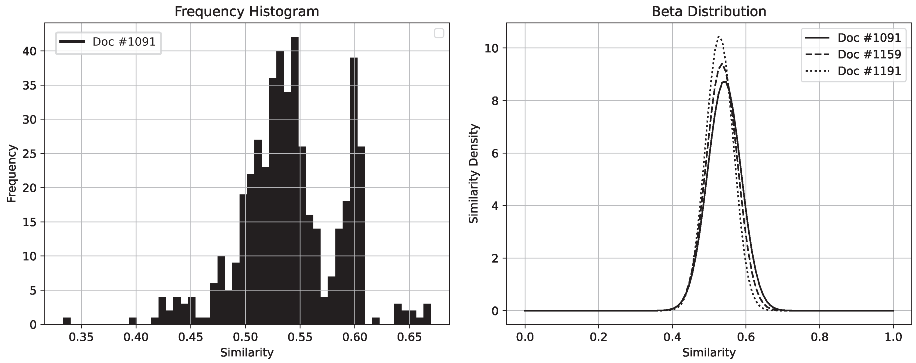

Fig. 5: The left shows the frequency histogram of continuous similarity values for one document. The right displays the beta distribution for three different documents constrained to binary values with δ and configuration A_2.

in the corpus consist of about 500 to 1,500 phrases. According to this, 1,000 time steps would meet the mean of the document length in the present corpus. Furthermore, this would lead to a net time budget of about 10 working days of eight hours. Which seems a reasonable effort for the research to create a sophisticated document.

The fully informed scenario calculated as explained in Section 3.4 is shown in Figure 4. It visualizes which strategy would be statistically optimal. The agent would then only use a selection of the best ranked documents with the highest mean. Figure 5 shows how the threshold turns the documents at random retrieval into a stochastic unit with a beta distribution of positive and negative action rewards. This distribution we interpret as the characteristic of the according bandit arms.

Epsilon Greedy Before applying complicated strategies we wanted to investigate how a simple mixture of greedy and random ratio would behave. We therefore let the epsilon greedy algorithm run with epsilon values from 0 to 1 in steps of 0.1. With $\varepsilon = 0$ having a complete greedy approach and $\varepsilon = 1$ having a fully random algorithm. We then compared the average reward of 10 epochs of 1000 simulations for the agent configurations A_1 and A_2. The graphical representation for a selection of epsilon values is shown in Figure 6.

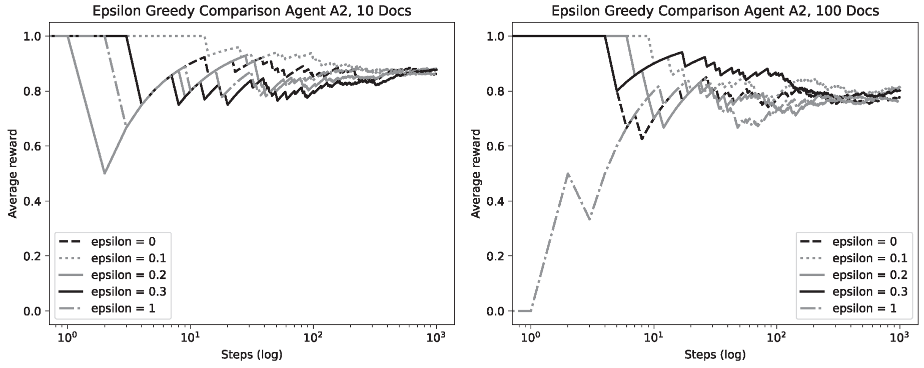

Fig. 6: Epsilon variation for 10 and 100 documents for agent configuration A_2.

Upper Confidence Bound The UCB approach was optimized but even so showed the worse performance of all algorithms. We see the reason for this in the high variance of the document similarity distribution. Additionally, the configuration of the hyper-parameters to determine the confidence interval and bias model requires efforts if applied to new corpora [3]. In a fixed scenario this might be reasonable and then the strategy could outperform simple EG. But in the present volatile scenario, UCB is not a policy to recommend.

Thompson Sampling The Thompson Sampling strategy incorporates the nature of the binary reward scenario. It models as well mean and variance of each document. It showed good results throughout the whole evaluation process and all configurations. It lacks in this basic version the possibility to adjust the ratio of exploration and exploitation.

Comparative Evaluation It turned out that the way of using a textual similarity to create a data-driven bandit model leads to a scenario of high variance. Combined with a constrained number of time steps this has significant influence on the solution policies. The EG approach has the advantage of being easily understandable and configurable. Additionally, the volatile scenario mitigates the shortcoming of revisiting "bad" arms. The high variance might also be the reason for the in total worst performance of the UCB approach as shown in Table 2. The Thompson algorithm showed good results and achieved in most experiments a higher reward in less time steps compared to the other algorithms. All algorithms are capable of approximately recreating the fully informed ranking of documents induced by each mean calculated in Figure 4. To pay respect to the aspects of configurability we suggest a combined approach of EG and TS. For instance, to randomly explore all arms for some steps and then start to use the Thompson Sampling approach.

Table 2. Overall performance of optimal, random, and bandit strategies.

Agent / Docs	A_1 **/ 10**	A_2 **/ 10**	A_1 **/ 100**	A_2 **/ 100**	**Mean Regret**
Optimal	0.68	0.93	0.69	0.89	0%
Random	0.53 (28%)	0.81 (19%)	0.54 (28%)	0.77 (16%)	22%
EG $\varepsilon = 0.2$	0.61 (11%)	**0.90 (3%)**	**0.59 (17%)**	**0.80 (11%)**	**11%**
UCB c=2	0.54 (26%)	0.83 (12%)	0.56 (23%)	0.78 (14%)	19%
TS	**0.62 (9%)**	0.88 (5%)	0.56 (23%)	0.79 (12%)	13%

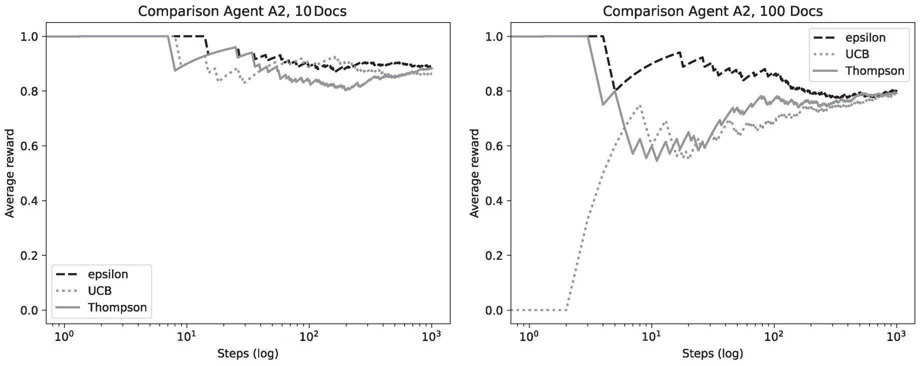

Fig. 7: Performance for 10 and 100 documents for agent configuration A_2.

5 Conclusion and Future Work

This work addressed the problem of distributing a limited budget of textual sampling over a corpus of documents. The lack of real life experts is mitigated by a data-driven agent model with special characteristics to estimate the quality of the samples for adaptational purposes. Therefore, we presented an approach to consider the documents of a corpus as bandit arms in a (constrained) multi-armed bandit setting. A case-based agent model was presented that can be configured together with a reward model to adjust the approach to different application scenarios. Hyper-parameters have been partially optimized in a learning phase. An evaluation was presented on a corpus of nuclear safety documents.

The evaluation showed that documents can be indeed considered as a kind of stochastically distributed entities with a mean reward under variation regarding a similarity-based reward model. The experiments showed that the characteristics of sample retrieval out of documents varies from a "real" stochastic bandit environment in severals aspects which leaves space for the following future work.

Several variations of the setting seem promising and other application domains are worth to investigate. This would give insight to the needs of other user groups with different requirements and possibly improve the performance of the present setting. The following aspects became obvious in the stage of development but eventually exceeded this work.

The underlying data has an inherent structure which can be exploited to create a more distinct bandit architecture. Even though the documents are discrete semantic objects, most likely, there will be a strong correlation amongst them if they belong to the same corpus. On the base of the correlation the reward could be adjusted for every bandit arm. The agent-teacher relationship could have an adversarial character. The teacher wants to challenge the agent. This could be implemented by reducing the reward of certain bandit arms.

An improvement of the agent model would be to switch from a stationary reward to a non-stationary setting that adapts the agent step by step to the character of the data. This could be used to learn agent models from the data. A further benefit would be to set up a contextual bandit model basing on this scenario to find correlations between the agent model and the data [16]. A distant perspective would be a full reinforcement learning model.

The calculation of textual similarities can be quite resource consuming. This might be an issue for complex similarity configurations of the agents, larger corpora, and applications with time pressure. In that case the complexity of algorithms will surely contain potential for improvement.

Finally, a case study with real life experts would help to refine the model. On the one hand it would be good to observe how real experts would choose an agent configuration according to their preferences and task characterization. A survey to determine hyper-parameters should yield insight into human behavior. For instance, taken the parameter δ which determines the threshold when to drop and when to keep a sample. At which similarity threshold does a real life expert tend to drop a sample?

References

1. International Atomic Energy Association: https://www.iaea.org
2. Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., Shamir, O.: Nonstochastic multi-armed bandits with graph-structured feedback. SIAM Journal on Computing **46** (09 2014)
3. Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research **3**, 397–422 (01 2002)
4. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.: The nonstochastic multiarmed bandit problem. SIAM Journal on Computing **32**, 48–77 (01 2003)
5. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence. pp. 2670–2676. IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007)
6. Bengs, V., Hüllermeier, E.: Multi-armed bandits with censored consumption of resources. Machine Learning **112**(1), 217–240 (2023)
7. Bergmann, R.: Experience Management. Springer, Berlin, Heidelberg (2002)
8. Brändle, F., Binz, M., Schulz, E.: Exploration Beyond Bandits, pp. 147–168. Cambridge University Press, Cambridge (2022)
9. Carlsson, E., Dubhashi, D.P., Johansson, F.D.: Thompson sampling for bandits with clustered arms. In: IJCAI International Joint Conference on Artificial Intelligence (2021)
10. Hartig, O.: Foundations of rdf-star and sparql-star (an alternative approach to statement-level metadata in rdf). In: Alberto Mendelzon Workshop on Foundations of Data Management (2017)
11. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics. pp. 539–545 (1992)
12. Hong, J., Kveton, B., Katariya, S., Zaheer, M., Ghavamzade, M.: Deep hierarchy in bandits. In: ICML International Conference on Machine Learning (2022)
13. Korger, A., Baumeister, J.: The SECCO ontology for the retrieval and generation of security concepts. In: Cox, M.T., Funk, P., Begum, S. (eds.) ICCBR. Lecture Notes in Computer Science, vol. 11156, pp. 186–201. Springer (2018)
14. Korger, A., Baumeister, J.: Case-based generation of regulatory documents and their semantic relatedness. In: Arei, K., Kapoor, S., Bhatia, R. (eds.) Future of Information and Communication Conference San Francisco. Advances in Information and Communication, vol. 1130, pp. 91–110. Springer (2020)
15. Kumar, S., Gao, H., Wang, C., Chang, K., Sundaram, H.: Hierarchical multi-armed bandits for discovering hidden populations. In: ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining. pp. 145–153 (08 2019)
16. Langford, J., Zhang, T.: The epoch-greedy algorithm for contextual multi-armed bandits. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. pp. 817–824. NIPS'07, Curran Associates Inc., Red Hook, NY, USA (2007)
17. Losada, D.E., Parapar, J., Barreiro, Á.: Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. Inf. Process. Manag. **53**, 1005–1025 (2017)

18. Perotto, F.S., Verstaevel, N., Trabelsi, I., Vercouter, L.: Combining bandits and lexical analysis for document retrieval in a juridical corpora. In: Bramer, M., Ellis, R. (eds.) Artificial Intelligence XXXVII. pp. 317–330. Springer International Publishing, Cham (2020)
19. Racharak, T., Suntisrivaraporn, B., Tojo, S.: sim-pi: A concept similarity measure under an agent's preferences in description logic elh. In: 8th International Conference on Agents and Artificial Intelligence. pp. 480–487 (01 2016)
20. Rahman, M.M., Kutlu, M., Lease, M.: Constructing test collections using multi-armed bandits and active learning. In: The Web Conference, San Francisco (05 2019)
21. Robbins, H.E.: Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society **58**, 527–535 (1952)
22. Schelling, T.C.: Dynamic models of segregation. The Journal of Mathematical Sociology **1**(2), 143–186 (1971)
23. Schulz, E., Franklin, N., Gershman, S.: Finding structure in multi-armed bandits. Cognitive Psychology **119** (10 2020)
24. Sen, R., Rakhlin, A., Ying, L., Kidambi, R., Foster, D., Hill, D., Dhillon, I.: Top-k extreme contextual bandits with arm hierarchy (02 2021)
25. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, second edn. (2018)
26. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**, 285–94 (1933)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
28. W3C: SKOS Simple Knowledge Organization System Reference: http://www.w3.org/TR/skos-reference (August 2009)
29. Wood, D., Lanthaler, M., Cyganiak, R.: RDF 1.1 concepts and abstract syntax (Feb 2014), http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/