 doi:10.20944/preprints202104.0575.v2

Communication

A Personalized Machine-Learning-enabled Method for Efficient Research in Ethnopharmacology. The case of Southern Balkans and Coastal zone of Asia Minor

Evangelos Axiotis [1,2]*, Andreas Kontogiannis [3], Eleftherios Kalpoutzakis [1] and George Giannakopoulos [4,5]

[1] Division of Pharmacognosy and Natural Products Chemistry, Department of Pharmacy, National and Kapodistrian University of Athens, axiotisevan@pharm.uoa.gr ; elkalp@pharm.uoa.gr
[2] Natural Products Research Center "*NatProAegean*", Gera, Lesvos, Greece.
[3] School of Electrical and Computer Engineering, National Technical University of Athens, Greece; andr.kontog@gmail.com
[4] Software and Knowledge Engineering Lab, NCSR "Demokritos", Athens, Greece, ggianna@iit.demokritos.gr
[5] SciFY PNPC, Greece

* Correspondence: axiotisevan@pharm.uoa.gr ; Tel.:00306944934227

Abstract: Ethnopharmacology experts face several challenges when identifying and retrieving documents and resources related to their scientific focus. The volume of sources that need to be monitored, the variety of formats utilized, the different quality of language use across sources, present some of what we call "big data" challenges in the analysis of this data. This study aims to understand if and how experts can be supported effectively through intelligent tools in the task of ethnopharmacological literature research. To this end, we utilize a real case study of ethnopharmacology research, aimed at the Southern Balkans and Coastal zone of Asia Minor. Thus, we propose a methodology for more efficient research in ethnopharmacology. Our work follows an "Expert-Apprentice" paradigm in an automatic URL extraction process, through crawling, where the apprentice is a Machine Learning (ML) algorithm, utilizing a combination of Active Learning (AL) and Reinforcement Learning (RL), and the Expert is the human researcher. ML-powered research improved 3.1 times the effectiveness and 5.14 times the efficiency of the domain expert, fetching a total number of 420 relevant ethnopharmacological documents in only 7 hours versus an estimated 36-hour human-expert effort. Therefore, utilizing Artificial Intelligence (AI) tools to support the researcher can boost the efficiency and effectiveness of the identification and retrieval of appropriate documents.

Keywords: Ethnopharmacology; Artificial Intelligence; Web Crawling; Active Learning; Reinforcement Learning; Text Mining; Big Data

1. Introduction

Ethnopharmacology is an interdisciplinary field of research based both on anthropological and scientific approaches [1]. The development of a standard scientific approach to retrieve information from the empirical use and define a pharmacological value from traditional preparations must be considered a highly complex and challenging task, strongly filtered by the evolution of human history [2].

In the Southern East European region, ethnobotanical studies are of great interest due to political and economic shifts that have influenced local lifeways, economies, foodways, and transmission of traditional knowledge regarding local health-related practices. [3].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202104.0575.v2

The challenge of discovering and enriching a body of knowledge with pre-existing scientific research has been a persistent need of the scientific community. Nowadays, intelligent systems known as "focused crawlers" [4], have supported domain experts in personalized search. Such approaches combine the power of the search engines with user's explicit feedback to identify the documents that maximally relate to the interest of the expert. The crawler leverages a limited set of keywords, provided by the users, to retrieve relevant documents. The experts, then, select the ones related to their interest and feed these back to the crawler. With subsequent iterations, the crawler can identify new keywords and fetch more pertinent documents by improving its searches.

Recent works employed data mining techniques to identify ethnopharmacology-related knowledge [5]. However, no work has yet provided personalized, adaptive, real-time support to experts. The present study focuses on the classification of the ethnopharmacological knowledge of Greece, southern Balkans, and the coastal zone of Asia Minor (Figure.1), with the broader aim to introduce a personalized computational approach to biomedical mining as an effective scientific tool for research in ethnopharmacology.

Figure 1. The zone of ethnopharmacological interest in white. Southern Balkans and coastal zone of Asia Minor.

This approach applies Machine Learning (ML) techniques, to get (a) automated inference on the explicit and implicit interests of the expert, (b) optimization of the crawling process to minimize the feedback of the expert on the appropriateness of retrieved documents. Our major contribution is that we propose an intelligent search system that practically supports the ethnopharmacological research through focused crawling, using a combination of Active Learning (AL) and Reinforcement Learning (RL).

2. Materials and Methods

2.1. Method Overview

Our work follows an "Expert-Apprentice" paradigm. The Expert has his/her personal interests and understanding of which publications actually relate to these interests. The Apprentice supports the Expert, by learning the interests in two ways. First, the Expert explicitly provides examples of documents, called "seeds". Second, over time the Apprentice periodically requests feedback from the Expert for an – ideally minimal - number of candidate documents. The Expert then labels them as interesting or not. The Apprentice resumes its work iteratively until it retrieves a specific number of documents.

In our Artificial Intelligence (AI) setting, as shown in the Flow diagram (Figure 2), we propose the Apprentice be an ML algorithm that undertakes 2 tasks. In the first task, the algorithm understands the interests of the user (Expert) through explicit feedback (labels of documents as interesting or not). Here, we utilize an ML model deploying pool-based AL for a binary classification task, with the Expert being the Oracle (human annotator) during the learning process. In the supervised pool-based AL setting, a model is trained on an initial small labeled training set of relevant and some irrelevant documents. Then, it queries the Oracle with the documents that are predicted to be the most informative for the model from a bigger unlabeled dataset, which is called "pool". After the Oracle has given the corresponding labels for these samples, the training set is augmented with them and the model retrained utilizing the updated data. This training process resumes iteratively until a predefined number of queries ("budget") has been addressed to the Oracle. We note that AL has already been used in other biomedical text mining applications [6,7], where classic ML classification algorithms, such as Support Vector Machine (SVM)[8] (a well-established classifier based on identifying representative instances that separate the classes of interest in a feature space), and Logistic Regression [9] (relying on a thresholded probability estimate mapping the input features of an instance to the probability of the instance to belong to each class) were examined. In our work, we utilize a common recurrent neural network, the Long-short term memory "LSTM" (a neural network embedding sequences to a vector space, making sure that similar sequences are positioned close to each-other in the embedding space) as the classification model for the AL setting.

In the second task, the Apprentice is an RL agent that discovers a strategy - policy - of crawling documents. The aim of the agent is to minimize the number of retrieved documents, while maximizing the number of relevant ones. To this end, the agent tries to connect the documents fetched so far with the decision of which candidate document to fetch next. We consider that we gather candidate documents from the references of each fetched publication. Every few fetched publications, the algorithm examines how well the strategy did in bringing relevant documents by using the trained AL model. The algorithm then updates its strategy, based on this feedback, trying to improve its decisions in future crawling steps. Thus, we utilize RL in order to optimize the automatic URL extraction process of focused crawler.

2.2. *Defining the Relevant Topics*

The relevant topics of our publication search are defined by the Expert. In our case, the relevant topics referred to ethnopharmacology in Balkan countries and Asia Minor with emphasis on certain plant families and species. More specifically, our domain experts pointed out 31 of the most important plant families. Using the taxonomy of angiosperms published on Flora of Greece [10], we managed to extract all species names from these families. Thus, we constructed a taxonomy of 578 keywords based on geographical locations and plant families.

2.3. *Dataset*

In the selected ethnopharmacology setting, we first examined whether two different researchers agree on the definition of relevance. This would imply that the topic of interest has been sufficiently described to gain common understanding between experts. To this end, we requested them to provide a list of 25 relevant documents – seeds [11] - identified by their URLs. Based on these seeds, we identified a total of 427 documents, which were extracted from the references of them.

We also retrieved another 800 publications, with no prior knowledge of whether they would be related to the topic at hand. This was achieved by a crawling run, which randomly followed references appearing in visited publications, through uniform sampling.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202104.0575.v2

By removing duplicates, we ended up with a total of 1012 documents, in addition to the seeds.

We arbitrarily selected a total of 50 documents, of which almost 50% were part of the seed set (very relevant). Then we asked independently the 2 domain experts to label the documents on a scale from 1 to 4 (1 = "highly related" and 4 = "irrelevant"). We then measured the degree of the inter-annotator agreement through three methods: Raw Agreement (RA) (counts the number of items for which the annotators provide identical labels), Cohen's kappa (CK) (takes into account the possibility of the agreement occurring by chance), and Krippendorff's alpha (KA) (measures the disagreement levels of annotators utilizing a distance function for each pair of labels) [12]. All methods showed substantial or good agreement between judges (RA: 0.82, CK: 0.71, KA: 0.92). This clearly showed that the experts do hold a common understanding of what is related to the domain of focus. Thus, the senior of the two experts undertook that annotation of data in the next experiments. The rate of annotation across experts was about 5 documents per minute, described only by their titles and abstracts. Thus, the annotation of the total 1012 documents by a single expert would have taken about 200 minutes. We note that this collection of documents would be the pool for our pool-based AL setting.

We now possess a means to obtain reference, agreed upon, opinions – referred to as "gold-standard" opinions on the relevance of a given document to our domain of interest. We can, thus, employ AL and crawling and evaluate how well the system (a) infers the interests of the expert(s) and (b) optimizes the crawling process to minimize the number of documents it needs to retrieve.

2.4. *Using Active Learning to Infer Expert Interest*

For the first aim, i.e., inferring what the expert considers related to the topic of interest, we trained an LSTM [13] model with AL, which implements part of the "Expert-Apprentice" workflow we described. Essentially, in our case, it refers to the algorithm which classifies a given document as relevant or not to the interest of the Expert. For this process, we set the budget of queries equal to 250, i.e., we can only ask the expert his/her opinion on a maximum of 250 documents. The document pool consists of the 1012 unlabeled documents collected using the random crawling run and those extracted from the seeds.

For reproducibility purposes we briefly describe our LSTM network which takes as input a sequence of pre-trained word2vec word embeddings of each document, based on the bio.nlplab.org embedding [14]. The network uses a Mean Pooling layer to average the hidden state vectors of all timesteps, i.e., words in a document.[1]This layer is connected to two fully connected layers. The AL model selects from a pool those k documents for which the corresponding classification probabilities are the k smallest. In order our model to output probability values for each corresponding class, we use the Softmax as the activation function of the output layer. We arbitrarily use k = 10.

Next, we tried to understand if the system would help the expert to retrieve a sufficient number of related documents under a significantly reduced human time allocation. To this end, we ran 4-fold-cross-validation (4 experiments) [15]. In each AL experiment, the training set was initially composed of 23 relevant and 27 irrelevant documents, for a total of 50 documents. In each run, we kept 100 held-out documents, evaluating the performance of the AL prediction: 50 were related and 50 were not related to the topic at hand. We essentially asked the expert about 250 documents (vs. 1012 that he would have needed to evaluate if no active learning was employed), reducing the required time and

[1] More information about the concepts of neural networks, activation functions, different types of layers and hyperparameters can be found in [40]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202104.0575.v2

effort by approximately 75%. For this level of reduction, the AL model managed to classify correctly 88 out of 100 documents on average (88% accuracy).

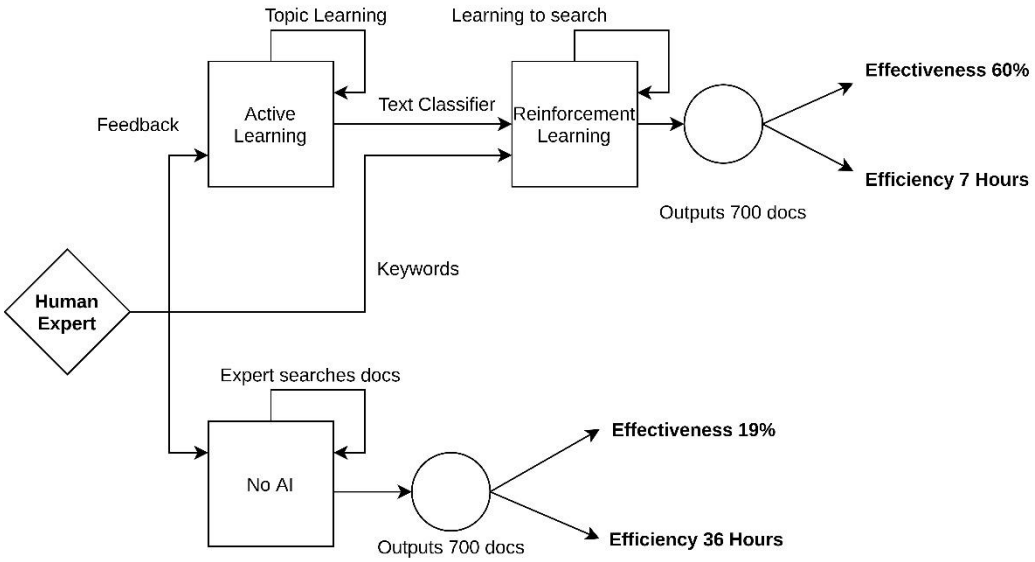

Figure 2. Flow diagram of the experimental method

2.5. Reinforcement Learning

In our setting, an RL algorithm allows the crawler to determine a strategy (policy), so that it retrieves a fixed number of documents while maximizing the number of related ones. Recently, there have been approaches of focused crawling [16] and biomedical data mining [17] with RL. An agent (the crawler) fetches URLs in an iterative manner. Each iteration is considered a timestep. The agent acts within a crawling environment. The environment has its state per timestep. There is a number of actions that the agent can take on each timestep. These actions lead to rewards over time. Formally, each timestep (t) the agent fetches a new URL, as a result of an action selection (A_t), then it transitions from the current state (S_t) to another state (S_{t+1}) and observes a reward (R_t). We consider the states to be related to the history of information (number of relevant and irrelevant URLs) fetched by the crawler. The actions are related to the URLs (keywords found on the anchor text) extracted from a state transition. The reward is related to the relevance of the current fetched publication with the defined topic. We set the reward equal to 1 for relevant publications and 0 otherwise. For the reward function, at first, we use the LSTM trained by AL in order to decide whether a document is related to ethnopharmacology. Then, we deterministically filter the related predicted ones by using the taxonomy of keywords constructed.

The goal of the agent is to find a policy (utilizing an RL algorithm), to maximize the discounted cumulative received reward $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots + \gamma^{T-t} R_T$, [18] where T is the fixed number of total documents that the crawler should fetch and γ is the discount factor. In other words, the agent seeks to find a mapping between states and actions, in order to get high long-term rewards. For our experiment, we arbitrarily set T = 700 and γ = 0.99.

Our evaluation measure for focused crawling is the harvest rate HR(t) [4], which is the cumulative percentage of relevant fetched documents up to timestep t. Formally, it is defined as

$$HR(t) = \frac{Number\ of\ Relevant\ documents\ fetched\ since\ t}{Number\ of\ all\ documents\ fetched\ since\ t}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202104.0575.v2

Since the RL agent is used to optimize the automatic URL extraction process and taking into account that the reward is 1 when the fetched webpage is relevant to our topic, harvest rate is also an evaluation measure for RL. It actually measures the mean cumulative reward that agent receives during the whole learning (crawling) process. Thus optimizing harvest rate is always equal to optimizing the mean cumulative reward of the RL agent.

We employ a Deep Q-learning approach, utilizing the Deep Q-Network (DQN) agent [19], which is based on the TD Error [18], $R_{t+1} + \max_a Q^{\pi'}(S_{t+1}, a; \theta^-) - Q^{\pi}(S_t, A_t; \theta)$, where Q^{π} and $Q^{\pi'}$ are the action-value functions under the policies π and π', respectively. That is $Q^{\pi}(S_t, A_t) = E_{U(D)}[R_{t+1} + \max_a Q^{\pi'}(S_{t+1}, a; \theta^-) \mid S_t, A_t]$. Which reflects the expected cumulative (long-term) rewards given current state S_t, current action A_t and immediate reward R_{t+1} .The DQN agent consists of two neural networks with the same architecture - a Q-Network (θ) and a Target Q-Network (θ^-) - in order to approximate Q^{π} and $Q^{\pi'}$, respectively. Additionally, it has a replay buffer D, called Experience Replay, which is important for uniform sampling mini-batches of uncorrelated past state transitions. For each Q-Network, we utilize a Multilayer perceptron (MLP) with two hidden layers. We initialize the Experience Replay with a priori experience given from seeds, all of which are highly relevant documents, in order to speed up the training process. Using Deep Q-learning, we essentially face a regression problem, minimizing the Mean Square Error of TD Error with respect to θ. Moreover, to balance the Exploration-Exploitation dilemma, calling us to decide between always choosing the best action (exploiting) and uniformly selecting sometimes one (exploring), we use an ε-greedy policy for sampling, i.e., action selection. That is, the best action of a given state is chosen with probability 1-ε, otherwise a random one is selected (with probability ε). As training progresses, ε diminishes over time by a factor of λ until it reaches a defined value εF. Formally, $\varepsilon = \max\{\varepsilon F, \lambda\varepsilon\}$. We set $\lambda = 0.99$, initial $\varepsilon_0 = 0.15$ and $\varepsilon F = 0.03$.

For our agent to be able to select URLs - related to actions - extracted from past state transitions, we use a priority queue, called the frontier, so that the best action is selected in O(log(N)); where N is the frontier size. We note that a URL is stored into the frontier along with its corresponding Q-value, which was estimated by the Q-Network. Also, we define another structure, called closure which represents a utility structure, essentially a map/dictionary (essentially a set of key-value pairs). There we store fetched URLs, so that the agent will not fetch them again.

Finally, we can describe the proposed focused crawling process that our agent follows. At this point, we consider that the AL process has been completed. Thus, we have a trained LSTM model for predicting whether a document (publication) is relevant to our topic of interest. Recall that the predictions of this model are first filtered, using a given taxonomy of keywords, in order to give the corresponding rewards that the agent receives during the whole crawling process. At first, user gives a few seed references (URLs), which are all highly relevant to the topic of interest, along with the taxonomy of keywords. These seeds are the starting point of the crawling process. As we mentioned above, the corresponding information from them is stored in the Experience Replay, before the crawling process starts. Also, the references extracted from the seed publications are stored in frontier with an initial Q-value, while the seed URLs are saved in closure. Recall that we use the closure structure in order not to fetch a URL more than once.

When the crawling process has started, at each timestep the DQN agent, given its state, samples an action (related to a URL) from the frontier using the ε-greedy policy. After fetching the corresponding publication, its references are extracted and stored in frontier along with a corresponding Q-value computed by the agent. At the same time, the URL of the fetched publication is stored in closure. Selecting an action from frontier, the agent then receives a reward. Then, it transitions to another state, related to the current fetched publication and the history of publications fetched during the whole crawling

 doi:10.20944/preprints202104.0575.v2

process. This state transition is then stored in the Experience Replay. Then, the agent learns from the past transitions, according to the Deep Q-learning algorithm. Note that this procedure is repeated iteratively, until a predefined number of publications is fetched by the focused crawler.

We note that for the training of the above neural network, we use Adam optimizer with initial learning rate equal to 0.001. Also, for each training step, we sample from Experience Replay with constant batch size equal to 16. We set the target update period equal to 100, that is the weight values of the Q-Network are copied to the Target Q-Network after 100 (crawling) timesteps. Thus, during our total 700 crawling timesteps process, the Target Q-Network is updated 7 times. Moreover, in order to collect some more data, our agent starts learning after the 40 timesteps have passed. We note that for these 40 timesteps, we perform only exploration utilizing random crawling, i.e. a URL is selected from the frontier with uniform sampling.

Last but not least, at this point we will discuss some more implementation details. We developed our focused crawler system using Python 3 [**20**]. More specifically, we used Keras [**21**] and TensorFlow 2 [**22**] for building and training all neural networks described in Sections 2.4 and 2.5. Also, we built the crawling environment utilizing the open-source toolkit Gym [**23**]. We note that the whole crawling process was conducted using URLs from PubMed [**24**] and MEDLINE [**25**]. For this aim, in order to fetch webpages and access reference publication, we utilized the open-source tool PubMed_parser [**26**].

3. Results

3.1. Ethnopharmacological Inference

Ethnobotany in the Southern East (SE) European region includes local traditional knowledge from countries such as Albania [27], FYROM [28], Bulgaria [29], and Greece [30,31,32]. In the present study, the coastal zone of Asia Minor is included [33,34,35]. The conspicuous floristic affinities of the East Aegean islands with neighboring western Anatolia, along with the enduring influence that Anatolian Turks had on eastern Europe during the Ottoman empire, prompted us to compare data of ethnopharmacological studies from this area.

The Balkan area can be described both as a “linking bridge” of cultures and as a violent transitional zone between civilizations; the bio-cultural-historical amalgam of races in the southern part of the peninsula represents the core of “Balkanization” [36], a concept coined to define the anthropological mixture in SE.

Moving towards the southern parts of the peninsula, a unique cultural and linguistic pattern has evolved with populations influenced by the dominance of ancient Macedonians (500-168 BC), Romans (168-284 BC), Byzantines (395-1453 AD), and Ottomans (1299-1922 AD). From the beginning of the 19th century, the Balkans were transformed from protectorates of foreign empires to independent countries, but the cultural amalgam was so intertwined that was embodied by the borders of these nation-states even after many generations. Even if hundreds of different ethnic groups exist in these countries, they are incorporated into the local societies in such a way that it is very difficult to investigate their origin [37]. In many instances, researchers described an erosion of traditional medical knowledge due to deep social changes [3]. As a result, the loss of information is inevitable.

Moreover, rich biodiversity characterizes these regions and a great number of species have been used in traditional medicine. A non-exhaustive list of species in the earliest written records still preserved has been exploited by local healthcare [38].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202104.0575.v2

Lately, many online resources are trying to pass on this knowledge, mostly accounting for oral reports from elderly people. These attempts create a conspicuous variety of sources that needs new technologies to be processed [39], classified and validated, for the best advantage of the scientific community. In our project, we were faced with this great challenge. The volume of sources that needed to be monitored exceeded a database of 10,000 identified references, based on the topics summarized in Table 1. We limited the Plant Families in the classification of Angiosperms and from these we considered 31 of the most important plant families used in ethnopharmacology. Furthermore, the part of the plant used, uses and recipes, Medical Subject Headings (MeSH) terms, and geographical regions, were used to filter the identified references.

Table 1. Ethnopharmacological Topics / Mesh Terms used for our setting

Plant Families	Part of Plant used	Uses / Recipes	MeSH Terms	Geographical Regions
Alliaceae	Aerial Part	Decoction	Greek ethnopharmacology	Albania
Anacardiaceae	Flower	Infusion	Traditional greek medicine	FYROM or Northern Macedonia
Apiaceae	Chalices of flowers	Maceration	Natural product	Bulgaria (southern)
Asparagaceae	Seed	Powder	Medicinal plant	Greece
Asphodelaceae	Leaf	Juice	Plant extracts	coastal zone of Turkey or Asia Minor
Asteraceae	Fruit	Poultice	Pharmacological action	
Boraginaceae	Stem	tsp of oil	Disease	
Brassicaceae	Bark	Paste	Treatment	
Cactaceae	Root	Whole plant preparation	Antimicrobial activity	
Cannabaceae	Clove	Cook	Radical scavenging activity	
Capparaceae	Stigma	Raw	Antioxidant activity	
Cistaceae	Bulb	Milk	Ethnobotany	
Fabaceae	Foliage	Solvent / adjuvant used	Pharmacognosy	
Fagaceae	Shoot	Honey	Herbal medicine	
Gentianaceae	Branch	Wine / Water	Greek folk medicine	
Hypericaceae	Whole Plant	Filtrate	Home remedies	
Lamiaceae	Wooden	Pounded	Folk remedies	
Liliaceae	Kernel	Extract	Materia medica	
Malvaceae	Fiber	Dried	Phytotherapy	
Moraceae	Rhizome	Fresh	Southern Balkans	
Myrtaceae	Ground plant	Soup	Balkans	
Oleaceae	Petioles	Soaked in	Albanian ethnopharmacology	
Paeoniaceae	Stem bark	Milled	Bulgarian ethnopharmacology	
Platanaceae	Tuberous root	Mixed with	Southern Bulgary ethnopharmacology	
Rosaceae	Styles	Warm and smoke	FYROM ethnopharmacology	
Salicaceae	Latex	Chew	Northern Macedonia ethnopharmacology	
Scrophulariaceae	Gum	Swallow	Turkish ethnopharmacology	
Solanaceae	Peels	Bake	Turkish coastal zone ethnopharmacology	
Urticaceae	Ripe ears	Bandage	Turkish folk medicine	
Valerianaceae	Hard wood	Squeez	Southern Balkans folk medicine	
Vitaceae	Radix	Disperse in	Pharmacotherapy	

3.2. *Crawling Results*

In a baseline setting, automatic crawling would just exhaustively return the references of the seeds, and then recursively the references of these references. This causes a significant growth in the number of fetched documents, without ascertaining quality results. A human, on the other hand, would follow a much more targeted approach, by evaluating the most promising documents each time, visiting them, and in turn, judging their references. In the RL setting, the agent may determine that in some cases it is promising to follow a marginally relevant reference, to then reach a wealth of other publications that might have not been fetched with the previous method.

In this case, we measure the reduction in crawled publications, compared to the baseline. We also take into account how many documents retrieved were indeed relevant to our topic. We note that in the baseline approach:

- in the first 25 documents, we have approximately 850 references to visit;

- in the first 700 fetched documents, the identified references are approximately 10,000.

We have estimated, by sampling 50 representative documents, that the percentage of related references per document is approximately 19%. On the other hand, our DQN agent fetched 700 documents, measuring the HR as 60% (420 relevant documents from 700), i.e., improving 3.1 times the effectiveness over the baseline. Recall that this HR score

is also the mean cumulative reward the agent received during the whole crawling (learning) process.

As a second aspect, we examined the same number (420) of related documents the expert can retrieve in the unit of time. Taking into account the time needed for the expert to annotate a single document, we estimate that they need a total time of 36 hours for this task, which is a rate of 13 relevant documents per hour. The RL-based system achieved a rate of 68 relevant documents per hour through a 7-hour crawling task, and thus improved 5.14 times the efficiency over the expert.

5. Conclusions

In this study, we demonstrated a methodology utilizing AL and RL methods that can significantly boost the effectiveness and efficiency of ethnopharmacology researchers. Moreover, we demonstrated that AI-powered research can improve 3.1 times the effectiveness and 5.14 times the efficiency of the domain expert, suggesting the use of such tools for ethnopharmacology research. After this preliminary study, we can safely hypothesize that the use of AI tools can indeed support the researchers to boost the efficiency and effectiveness of the identification and retrieval of appropriate documents. For future work, we plan to develop a streamlined end-to-end software system, combining the developed (back-end) methodology with an intuitive (front-end) user experience, practically supporting ethnopharmacological research workflows. The contribution of this system to everyday practice would be the significant reduction of time and effort allocated to the identification and collection of documents relevant to a researcher's focus.

Author Contributions: Conceptualization, E.A., A.K. and G.G.; methodology, E.A., A.K., G.G; software, A.K. and G.G.; validation, E.A., A.K. and G.G.; formal analysis, E.A, A.K. and G.G.; investigation, E.A., A.K., G.G.; resources, E.A., A.K., G.G. and E.K..; data curation, E.A., A.K. and G.G.; writing—original draft preparation, E.A., A.K. and G.G..; writing—review and editing, E.A.,A.K. and G.G.; visualization, E.A.,A.K.,G.G. and E.K.; supervision, E.A. and G.G.; project administration, E.A. and G.G.; funding acquisition, E.A. and E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Heinrich, M.; Jäger, A. K. *Ethnopharmacology*, John Wiley & Sons: Chichester, UK, England, 2015
2. Lukman, S.; He, Y.; Hui, S.C. Computational methods for Traditional Chinese Medicine: A survey. *Comput. Methods Programs Biomed.* **2007**, 88, pp. 283–294. https://doi.org/10.1016/j.cmpb.2007.09.008
3. Quave, C.L.; Pardo-De-Santayana, M.; Pieroni, A. Medical ethnobotany in Europe: From field ethnography to a more culturally sensitive evidence-based cam? *Evid-Based Compl. Alt.* **2012**. https://doi.org/10.1155/2012/156846
4. Chakrabarti, S.; Van den Berg, M.; Dom, B. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer networks* **1999**, 31, pp. 1623–1640.
5. Yadong, Z.; Kongfa, H.; Tao, Y. Mining effect of Famous Chinese Medicine Doctors on Lung-cancer based on Association rules. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2019**, pp. 2036–2040.
6. Naseem, U.; Khushi, M.; Khan, S.K.; Shaukat, K.; Moni, M.A. A Comparative Analysis of Active Learning for Biomedical Text Mining. *Appl. Syst. Innov.* **2021**, 4, 23. https://doi.org/10.3390/asi4010023
7. Chen, Y.; Mani, S.; Xu, H. Applying active learning to assertion classification of concepts in clinical text. *J. Biomed. Inform.* **2012**, 45, pp. 265–272. https://doi.org/10.1016/j.jbi.2011.11.003
8. Cortes, C., Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, 20, pp.273–297. https://doi.org/10.1007/BF00994018
9. McCullagh, P., Nelder, J. A. *Generalized Linear Models*. London: Chapman & Hall / CRC, UK, England, 1989.
10. Flora of Greece. Vascular Plant Checklist of Greece. Available online: http://portal.cybertaxonomy.org/flora-greece/ (accessed on 16 June 2021)
11. GitLab Repository. Available online: https://gitlab.com/andr_kontog/seed_urls/-/blob/main/seeds_25.txt (accessed on 16 June 2021).

12. Arstein R. Inter-annotator Agreement. In *Handbook of Linguistic Annotation*; Pustejovsky J., Eds.; Springer, Dordrecht, Holland, 2017. https://doi.org/10.1007/978-94-024-0881-2_11
13. Hochreiter, S., Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, 9, (8), pp. 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735
14. Biomedical natural language processing. Tools and resources. Available online: https://bio.nlplab.org/ (accessed on 16 June 2021)
15. Hastie, T., Tibshirani, R., & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed.; Springer: New York, USA, 2009.
16. Han M.; Wuillemin PH.; Senellart P. Focused Crawling Through Reinforcement Learning. In *Web Engineering. ICWE. Lecture Notes in Computer Science*; Mikkonen, T., Klamma, R., Hernández J., Eds.; Springer, Cham., 2018, Vol 10845. https://doi.org/10.1007/978-3-319-91662-0_20
17. Souid, A.; Sakli, N.; Sakli, H. Classification and Predictions of Lung Diseases from Chest X-rays Using MobileNet V2. *Appl. Sci.* **2021**, 11, pp. 2751. https://doi.org/10.3390/app11062751
18. Sutton, R.S. & Barto, A.G. *Reinforcement learning: An introduction*; MIT press: USA, 2018.
19. Mnih, V.; Kavukcuoglu, K.; Silver, D. et al. Human-level control through deep reinforcement learning. *Nature* **2015**, 518, pp. 529–533. https://doi.org/10.1038/nature14236
20. Python 3. Available online: https://www.python.org/ (accessed on 16 June 2021).
21. Keras. Available online: https://keras.io/ (accessed on 16 June 2021).
22. TensorFlow 2. Available online: https://www.tensorflow.org/ (accessed on 16 June 2021).
23. Gym. Available online: http://gym.openai.com/ (accessed on 16 June 2021).
24. PubMed. Available online: https://pubmed.ncbi.nlm.nih.gov/ (accessed on 16 June 2021).
25. MEDLINE. Available online: https://www.nlm.nih.gov/medline/medline_overview.html (accessed on 16 June 2021).
26. Titpata Python Parser. Available online: https://github.com/titipata/pubmed_parser (accessed on 16 June 2021).
27. Pieroni, A. Local plant resources in the ethnobotany of Theth, a village in the Northern Albanian Alps. *Genet. Resour. Crop Evol.* **2008**, 55, pp.1197–1214. https://doi.org/10.1007/s10722-008-9320-3
28. Miskoska-Milevska, E.; Stamatoska, A.; Jordanovska, S. Traditional uses of wild edible plants in the Republic of North Macedonia. *Phytol. Balc*. **2020**, 26, pp.155–162.
29. Ivanova, T.A.; Bosseva, Y.Z.; Ganeva-Raycheva, V.G.; Dimitrova, D. Ethnobotanical knowledge on edible plants used in zelnik pastries from Haskovo province (Southeast Bulgaria). *Phytol. Balc.* **2018**, 24, pp. 389–395.
30. Vokou, D.; Katradi, K.; Kokkini, S. Ethnobotanical survey of Zagori (Epirus, Greece), a renowned centre of folk medicine in the past. *J. Ethnopharmacol.* **1993**, 39, pp. 187–196. https://doi.org/10.1016/0378-8741(93)90035-4
31. Axiotis, E.; Halabalaki, M.; Skaltsounis, L.A. An ethnobotanical study of medicinal plants in the Greek islands of North Aegean Region. *Front. Pharmacol.* **2018**,9, pp.1–6. https://doi.org/10.3389/fphar.2018.00409
32. Tsioutsiou, E.E.; Giordani, P.; Hanlidou, E.; Biagi, M.; De Feo, V.; Cornara, L. Ethnobotanical Study of Medicinal Plants Used in Central Macedonia, Greece. *Evidence-based Complement. Altern. Med.* **2019**. https://doi.org/10.1155/2019/4513792
33. Ugulu, I.; Baslar, S.; Yorek, N.; Dogan, Y. The investigation and quantitative ethnobotanical evaluation of medicinal plants used around Izmir province, Turkey. *J. Med. Plants Res.* **2009**, 3, pp. 345–367. https://doi.org/10.5897/JMPR.9001216
34. Kargıoğlu, M.; Cenkci, S.; Serteser, A.; Konuk, M.; Vural, G., Traditional uses of wild plants in the middle Aegean region of Turkey. *Hum. Ecol.* **2010**, 38, pp.429–450. https://doi.org/10.1007/s10745-010-9318-2
35. Polat, R.; Satil, F. An ethnobotanical survey of medicinal plants in Edremit Gulf (Balikesir - Turkey). *J. Ethnopharmacol.* **2012**,139, pp. 626–641. https://doi.org/10.1016/j.jep.2011.12.004
36. Ballinger, P. Definition Dilemmas: Southeastern Europe as a «Culture Area»? *Balkanologie* **1999**, 2. https://doi.org/10.4000/balkanologie
37. Carter, F.W. *An Historical Geography of the Balkans*; Academic Press, NY, USA, 1977; pp.580.
38. Legakis, A.; Constantinidis, T.; Petrakis, P. V. Biodiversity in Greece, *Global Biodiversity* **2018**. https://doi.org/10.1201/9780429487750-4
39. Yao, Y.; Wang, Z.; Li, L.; Lu, K.; Liu, R.; Liu, Z.; Yan, J. An Ontology-Based Artificial Intelligence Model for Medicine Side-Effect Prediction: Taking Traditional Chinese Medicine as an Example. *Comput. Math. Methods Med*. **2019**. https://doi.org/10.1155/2019/8617503
40. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: USA, 2016.