A Personalized Machine-Learning-Enabled Method for Efficient Research in Ethnopharmacology. The Case of the Southern Balkans and the Coastal Zone of Asia Minor

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

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, and 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 the 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 the effectiveness and efficiency of the domain expert by 3.1 and 5.14 times, respectively, fetching a total number of 420 relevant ethnopharmacological documents in only 7 h versus an estimated 36 h of 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 on both anthropological and scientific approaches [1]. The development of a standard scientific approach to retrieve information from empirical use and define a pharmacological value from traditional preparations is considered a highly complex and challenging task, strongly filtered by the evolution of human history [2].

In the southeastern 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].

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], support domain experts in personalized searches. Such approaches combine the power of search engines with the user’s explicit feedback to identify the documents that maximally relate to the interest of the expert.
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 have 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 ethnopharmacological knowledge of Greece, the southern Balkans, and the coastal zone of Asia Minor (Figure 1), with the broader aim of introducing 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.](image)

This approach applies machine learning (ML) techniques to get (a) automated inference on the explicit and implicit interests of the expert and (b) optimization of the crawling process to minimize the feedback of the expert on the appropriateness of the retrieved documents. Our major contribution is that we propose an intelligent search system that practically supports 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 in Figure 2, we propose the apprentice be an ML algorithm that undertakes two tasks. In the first task, the algorithm understands the interests of the user (expert) through explicit feedback (the labeling 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 a supervised pool-based AL setting, a model is trained on an initial small, labeled training set of relevant and 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 a “pool”. After the oracle has given the corresponding labels for these samples, the training set is augmented with them, and the model is 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 belonging to each class) have been examined. In our work, we utilize a common recurrent neural network, “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 is doing in retrieving 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 the 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 refer 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 in 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.

Figure 2. Flow diagram of the experimental method.
2.3. Dataset

In the selected ethnopharmacology setting, we first examined whether two different researchers would agree on the definition of relevance. This would imply that the topic of interest has been sufficiently described to gain a common understanding between experts. To this end, we requested they 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 lists of references in 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 the visited publications through uniform sampling. 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 the 2 domain experts to independently label the documents on a scale from 1 to 4 (1 = “highly related” and 4 = “irrelevant”). We then measured the degree of 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 the annotators utilizing a distance function for each pair of labels) [12]. All methods showed substantial or good agreement between the judges (RA: 0.82, CK: 0.71, KA: 0.92). This clearly showed that the experts held a common understanding of what is related to the domain of focus. Thus, the senior of the two experts undertook the annotation of data in the next experiment. The rate of annotation across experts was about 5 documents per minute, described only by their titles and abstracts. Thus, the annotation of all 1012 documents by a single expert would have taken about 200 min. We noted 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 have described. Essentially, in our case, it refers to the algorithm that 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 will briefly describe our LSTM network, which takes as input a sequence of pretrained 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. This layer is connected to two fully connected layers (more information about the concepts of neural networks, activation functions, different types of layers, and hyperparameters can be found in [15]). The AL model selects from a pool those k documents for which the corresponding classification probabilities are the k smallest. In order for our model to output probability values for each corresponding class, we use 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 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) [16]. 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 effort by approximately 75%. For this level of reduction, the AL model managed to classify 88 out of 100 documents correctly, on average (88% accuracy).

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 [17] and biomedical data mining [18] 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 at each timestep. These actions lead to rewards over time. Formally, at 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 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 $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots + \gamma^{T-t} R_T$ [19], where $T$ is the fixed number of total documents that the crawler should fetch and $\gamma$ 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 $\gamma = 0.99$.

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

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

Since the RL agent is used to optimize the automatic URL extraction process, taking into account that the reward is 1 when the fetched webpage is relevant to our topic, the harvest rate is also an evaluation measure for RL. It actually measures the mean cumulative reward that the agent receives during the whole learning (crawling) process. Thus, optimizing the 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 [20], which is based on the TD error [19], $R_{t+1} + \max_a Q^\pi(S_{t+1}, a; \theta) - Q^\pi(S_t, A_t; \theta)$, where $Q^\pi$ and $Q'^\pi$ are the action-value functions under the policies $\pi$ and $\pi'$, respectively. That is, $Q^\pi(S_t, A_t) = E[U(D) | R_{t+1} + \max_a Q'^\pi(S_{t+1}, a; \theta^-) | S_t, A_t]$. This 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 ($\theta$) and a target Q-Network ($\theta^-$)—in order to approximate $Q^\pi$ and $Q'^\pi$, respectively. Additionally, it has a replay buffer, $D$, called experience replay, which is important for the uniform sampling of 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 $\theta$. Moreover, to balance the exploration–exploitation dilemma, which requires us to decide between always choosing the best action (exploiting) and, sometimes, uniformly selecting one (exploring), we use an $\varepsilon$-greedy policy for sampling, i.e., action selection. That is, the best action of a given state is chosen with probability $1-\varepsilon$; otherwise, a random one is selected (with probability $\varepsilon$). As training progresses, $\varepsilon$ diminishes over time by a factor of $\lambda$ until it reaches a defined value $\varepsilon_F$. 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. Additionally, we define another structure, called closure, that 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, the 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. Additionally, the references extracted from the seed publications are stored in the 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 starts, at each timestep, the DQN agent, given its state, samples an action (related to a URL) from the frontier using the $\varepsilon$-greedy policy. After fetching the corresponding publication, its references are extracted and stored in the 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 the 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 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 used the Adam optimizer with an initial learning rate equal to 0.001. Additionally, for each training step, we sampled from experience replay with a 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 the entire 700 crawling timesteps process, the target Q-Network is updated 7 times. Moreover, in order to collect more data, our agent starts learning after 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 more implementation details. We developed our focused crawler system using Python 3 [21]. More specifically, we used Keras [22] and TensorFlow 2 [23] for building and training all neural networks described in Sections 2.4 and 2.5. Additionally, we built the crawling environment utilizing the open-source toolkit Gym [24]. We note that the whole crawling process was conducted using URLs from PubMed [25] and MEDLINE [26]. For this aim, in order to retrieve webpages and access reference publications, we utilized the open-source tool PubMed_parser [27].
3. Results

3.1. Ethnopharmacological Inference

Ethnobotany in the southeastern (SE) European region includes local traditional knowledge from countries such as Albania [28], Republic of North Macedonia [29], Bulgaria [30], and Greece [31–33]. In the present study, the coastal zone of Asia Minor is included [34–36]. The conspicuous floristic affinities of the East Aegean islands with neighboring western Anatolia, along with the enduring influence that Anatolian Turks have had on eastern Europe during the Ottoman empire, prompted us to compare the data of ethnopharmacological studies from this area.

The Balkan area can be described as both a “linking bridge” of cultures and a violent transitional zone between civilizations; the biocultural–historical amalgam of races in the southern part of the peninsula represents the core of “Balkanization” [37], a concept coined to define the anthropological mixture in the SE.

Moving towards the southern parts of the peninsula, a unique cultural and linguistic pattern has evolved with the 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 into independent countries, but the cultural amalgam was so intertwined that it was embodied in 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 [38]. In many instances, researchers have described an erosion of traditional medical knowledge due to great 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 systems [39].

Lately, many online resources have tried to pass on this knowledge, mostly oral reports from elderly people. These attempts create a conspicuous variety of sources that needs new technologies in order to be processed [40], classified, and validated for the 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 to 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 heading (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 |
### Table 1. Cont.

| Plant Families | Part of Plant Used | Uses/Recipes | MeSH Terms                  | Geographical Regions                  |
|----------------|--------------------|--------------|-----------------------------|----------------------------------------|
| Asteraceae     | Fruit              | Poultice     | Pharmacological action      |                                        |
| Boraginaceae   | Stem               | tsp of oil   | Disease                     |                                        |
| Brassicaceae   | Bark               | Paste        | Treatment                   |                                        |
| Cactaceae      | Root               | Whole plant  | 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 Bulgarian 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 |                                  |

### 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 the quality of the 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 not have been retrieved 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 had approximately 850 references to visit;
- in the first 700 fetched documents, the identified references were 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 retrieved 700 documents, with the HR measuring 60% (420 relevant documents from 700), i.e., improving the effectiveness over the baseline 3.1 times. Recall that this HR score is also the mean cumulative reward the agent received during the entire crawling (learning) process.

As a second aspect, we examined the same number (420) of related documents the expert can retrieve in a unit of time. Taking into account the time needed for the expert to annotate a single document, we estimate that they will need a total time of 36 h 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 h crawling task and, thus, has the improved efficiency over the expert of 5.14 times.

4. Conclusions

In this study, we have demonstrated a methodology utilizing AL and RL methods that can significantly boost the effectiveness and efficiency of ethnopharmacology researchers. Moreover, we have demonstrated that AI-powered research can improve the effectiveness and efficiency of the domain expert by 3.1 and 5.14 times, respectively, 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 researchers by boosting 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 to practically support 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. and 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. and 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [https://gitlab.com/andr_kontog/seed_urls/-/blob/main/seeds_25.txt/, accessed on 16 June 2021].

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

References

1. Heinrich, M.; Jäger, A.K. *Ethnopharmacology*; John Wiley & Sons: Chichester, UK, 2015.
2. Lukman, S.; He, Y.; Hui, S.-C. Computational methods for Traditional Chinese Medicine: A survey. *Comput. Methods Programs Biomed.* 2007, 88, 283–294. [CrossRef] [PubMed]
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 Complement. Altern. Med.* 2012, 2012, 156846. [CrossRef] [PubMed]
4. Chakrabarti, S.; Van den Berg, M.; Dom, B. Focused crawling: A new approach to topic-specific Web resource discovery. *Comput. Netw.* 1999, 31, 1623–1640. [CrossRef]
5. Yadong, Z.; Kongfa, H.; Tao, Y. Mining effect of Famous Chinese Medicine Doctors on Lung-cancer based on Association rules. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 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. [CrossRef]
39. Legakis, A.; Constantinidis, T.; Petrakis, P.V. Biodiversity in Greece. In *Global Biodiversity*; Apple Academic Press: Boca Raton, FL, USA, 2018. [CrossRef]

40. 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, 2019, 8617503. [CrossRef] [PubMed]