Semantic Knowledge Extraction for Plant Species Diversity (Postprint)

Authors: Liu Jianhua, Wang Ying, Zhang Zhixiong, Li Chuanxi

Date: 2017-11-08T00:00:00+00:00

Abstract

Objective: To extend the species-centered plant species diversity extraction framework and explore the implementation of semantic knowledge extraction methods. Methods: Integrating current mainstream research on biodiversity extraction, we design a species-centered knowledge extraction framework encompassing multiple entities and their interrelationships, and leverage existing specialized databases to design and implement corresponding recognition methods. Results: We design a species-centered knowledge extraction framework, explore and implement semantic knowledge extraction methods for multiple entities and their interrelationships, and extend the extraction content and approaches in the field of plant species diversity. Limitations: The completeness and accuracy of entity recognition are heavily influenced by the underlying knowledge base, and the types of inter-entity relationships are limited to several categories such as co-occurrence, hierarchical, and syntactic relationships, requiring further investigation. Conclusion: This study extends the extraction content and approaches for plant species diversity, which can effectively support semantic retrieval and scientific computing.

Full Text

Preamble

Extracting Semantic Knowledge from Plant Species Diversity Collections

Liu Jianhua[1],[2], Wang Ying[1], Zhang Zhixiong[1], Li Chuanxi[3]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](China Great Wall Asset Management Co., Ltd, Beijing 100045, China)

Abstract

[**Objective**] This study aims to expand the species-centered extraction framework for plant species diversity and explore methods for semantic knowledge extraction. [**Methods**] Building upon mainstream biodiversity extraction research, we designed a knowledge extraction framework encompassing multiple entity types and inter-entity relationships, centering on species. Leveraging numerous existing specialized databases, we designed and implemented corresponding recognition methods. [**Results**] We developed a species-centric knowledge extraction framework and explored semantic knowledge extraction methods for various entities and their relationships, thereby expanding the scope and approach of plant species diversity extraction. [**Limitations**] The completeness and accuracy of entity recognition are significantly influenced by underlying knowledge bases, and relationship types are limited to co-occurrence, hierarchical, and syntactic relationships, requiring further investigation. [**Conclusions**] This research expands the content and methodology of plant species diversity extraction, effectively supporting semantic retrieval and scientific computation.

Keywords: Plant Species Diversity; Plant Species; Knowledge Extraction; Relation Extraction
Classification Number: G250

Introduction

Climate change, natural disasters, and other factors are accelerating species extinction at an unprecedented rate, making biodiversity conservation and sustainable utilization increasingly central to biodiversity research. Plant species constitute a critical component of biodiversity, attracting substantial research attention. A key challenge in plant species diversity informatics is helping researchers rapidly discover needed information from vast document collections rich in entities such as species names, genes, and experimental equipment. In response, researchers are actively working to leverage existing specialized plant species diversity databases—including species catalogs, herbarium collections, image libraries, and gene banks—to extract knowledge objects from descriptive texts and literature, enabling automatic deep indexing through semantic annotation techniques. This facilitates semantic integration and linking among digital resources, providing support for advanced semantic retrieval, data mining, and scientific computation.

Building upon existing research in plant species diversity information extraction and in line with the Chinese Academy of Sciences' National Science Library' s project requirements for "Constructing a Biodiversity Domain Ontology and Semantic Organization Application Demonstration Platform," this study targets plant species diversity as its domain. We designed a semantic knowledge organization framework for plant species diversity, explored extraction methods for semantic knowledge units defined within this framework, and developed a corresponding demonstration platform.

2. Related Research Overview

Through the efforts of many researchers, numerous information extraction tools for the biodiversity domain have emerged. These tools employ either single methods—such as natural language processing, dictionary matching, machine learning, rule-based templates, or shallow/deep syntactic parsing—or hybrid approaches combining several techniques. Most focus on recognizing various species names (scientific names, synonyms, common names, variant names, etc.), with some tools also addressing species trait identification. Thessen et al. [1] reviewed current research on species name recognition using natural language processing and machine learning algorithms in biodiversity science, while Naderi et al. [2] introduced various extraction tools for the biomedical domain under the GATE framework. These works provide comprehensive reviews of conventional biodiversity information extraction workflows, mainstream methods, and major tools at each stage. Rather than repeating these reviews, this paper focuses on key biodiversity information extraction tools to examine extraction content in the plant species diversity domain, providing a foundation for our proposed knowledge extraction framework.

Current plant species diversity extraction research can be summarized in three main areas:

2.1 Species Name Recognition and Standardization

Due to language variations and regional naming conventions, the same species name appears in scientific literature in diverse forms. Some follow the standard binomial (or trinomial) nomenclature in Latin, consisting of a genus name (capitalized) followed by a specific epithet (lowercase), both in full, often followed by the author abbreviation [3]. Others use abbreviated forms with only the genus initial and full specific epithet. Common names in English or other languages may also be used, with the same species potentially having different vernacular names across countries or regions [4]. These variations significantly increase recognition difficulty. Consequently, many researchers focus specifically on species name recognition, standardization, and organization, which represents the mainstream of current plant species diversity extraction research. Notable achievements include NCBI Taxonomy [5] and BioNames [6] (an online database linking animal names with source descriptions, classifications, and phylogenetic trees), the Species 2000 global species catalog [7], and mature recognition tools such as NetiNeti [8], OrganismTagger [9], Linnaeus [4], and TaxonGrab [10].

2.2 Species Trait Recognition

For taxonomic researchers, species trait descriptions—such as root, stem, and leaf color and length—are crucial for species classification. Some bioinformatics researchers have therefore explored automatic recognition methods for various species traits. Taylor [11] manually established rules and dictionaries based on

textual syntactic features to identify species parts, characteristics, and states. Tang et al. [12] built upon this work, using predefined templates for supervised learning to generate rules for recognizing leaf shape, size, color, arrangement, and fruit shape characteristics. CharaParser employs heuristic methods and syntactic feature-based rule generation to effectively identify multiple species traits [13]. Duan Yufeng et al. [14] continue to explore information extraction from Chinese plant species diversity descriptive texts.

2.3 Biological Network Recognition

Various biological entities (species, molecules, genes, proteins, etc.) exhibit multiple relationships that can be expressed as network graphs, enabling biological system analysis through graph-based methods [15-16]. Proteins and genes are focal points in biomedicine, with recognition research extending beyond plant species diversity. In plant species diversity literature, species phylogeny can be determined through gene sequencing, and proteins or genes can be used to influence or alter biological environments or characteristics. Therefore, recognizing proteins and genes involves not merely identifying named entities but extracting biological networks formed by various entities linked through verbs, prepositional phrases, possessives, and other syntactic structures.

Overall, current biodiversity information extraction research, particularly in plant species diversity, primarily focuses on exploring recognition methods for specific information types, aiming to structurally describe plant species diversity characteristics or assist in species identification. Systematic research and framework design for knowledge-based organization and semantic retrieval of scientific literature content remain rare. Based on existing research, this study systematically designs a semantic knowledge organization framework and explores practical methods for rapid identification of corresponding knowledge units.

3. Semantic Knowledge Framework Design

To conduct semantic knowledge extraction for plant species diversity, we must first identify what content to extract from target resources—that is, construct a reasonable semantic knowledge description framework. This framework serves as the foundation for describing semantic knowledge units and their relationships in the domain and supports subsequent knowledge organization and discovery. Therefore, based on analysis of existing research and project requirements for the "Biodiversity Domain Ontology and Semantic Organization Application Demonstration Platform," we designed a semantic knowledge framework to support this demonstration platform. This section details the framework' s design process and content.

3.1 Levels of the Semantic Knowledge Framework

During framework construction, we first searched the PubMed database using "Oryza sativa (rice species)" as the query term, targeting the journals *Plant Physiology* and *The Plant Cell*. From the retrieved collection, we randomly selected 100 scientific literature articles for manual annotation, with knowledge units confirmed by experts from the Institute of Botany, Chinese Academy of Sciences. Manual annotation proceeded at three levels, as illustrated in [Figure 1: see original paper].

[Figure 1: see original paper] shows the hierarchical example of manual annotation. (1) At the sentence level, the focus was on structuring scientific literature by identifying knowledge sentence groups serving specific purposes. In scientific literature, much knowledge cannot be simply represented as individual knowledge units (phrases) or inter-unit relationships. For example, a complete experimental condition (involving combined effects of chemical element concentrations and temperature control) or an entire experimental process may contain multiple knowledge units and relationships. For such information, we extract closely related phrases or short sentences to form knowledge sentence groups and determine their types (e.g., research methods, experimental procedures, research results), thereby reorganizing scientific literature knowledge. (2) At the knowledge unit level, we considered that scientific literature frequently contains numerous knowledge units with clear semantic categories, often appearing as named entities or phrases that convey deep textual content. Extracting these knowledge units enables fine-grained content revelation and is crucial for subsequent semantic retrieval. (3) At the knowledge unit relationship level, we recognized that knowledge units are not independent or scattered but form various semantic associations through co-occurrence, subject-predicate-object expressions, and other constructions. Combining these semantic associations maximizes deep text mining potential.

Among these three annotation levels, sentence-level extraction research is relatively independent and has been discussed in our previous work [17]; this paper therefore focuses on knowledge units and their interrelationships.

3.2 Content of the Semantic Knowledge Framework

In our semantic knowledge framework, knowledge units and their relationships constitute core content. Based on manual annotation results, current biodiversity extraction priorities, project requirements, and extraction feasibility, we designed the plant species diversity semantic knowledge framework shown in [Figure 2: see original paper]. The framework' s knowledge units (represented by boxes in [Figure 2: see original paper]) center on species and extend to various related knowledge units. For plant species attribute descriptions, we reused some concepts from the Plant Ontology (PO)[1], which essentially covers major knowledge points in current plant species diversity literature. Beyond hierarchical relationships (indicated by arrows in [Figure 2: see original paper]),

different knowledge unit categories also exhibit associations (non-arrow connections in [Figure 2: see original paper]). Through co-occurrence, syntactic, and semantic analysis, we can construct factual triples among these knowledge units to support further text analysis.

[1]The Plant Ontology, funded by the U.S. National Science Foundation (NSF), provides a controlled vocabulary for plant structure and growth stages.

4. Implementation of Semantic Knowledge Extraction

The plant species diversity semantic knowledge framework facilitates collecting, organizing, and structuring existing knowledge by storing records from plant ontology databases and other sources as instances of various knowledge types in [Figure 2: see original paper]. It also provides clear targets for further knowledge extraction. Based on this framework, we conducted knowledge extraction in the plant species diversity domain following the process described below.

4.1 Corpus Integration and Experimental Data Selection

After establishing the semantic knowledge organization framework, we compiled and integrated corpora through expert consultation and reference to relevant research from the Institute of Botany, Chinese Academy of Sciences [18]. This included the G2000 plant ontology database, NCBI Taxonomy [5], relevant domain terminology, geographical name vocabularies, and small compound names from Chemical Entities of Biological Interest[2]. We integrated instances according to the knowledge units defined in the semantic organization framework, ultimately compiling nearly 170,000 instance records. These domain resources serve both as direct lexicons for annotating knowledge unit instances and as bases for semi-automatically constructing entity recognition rule libraries to identify new instances.

Additionally, we obtained 23,000 journal abstracts from *Plant Physiology* and *The Plant Cell* in PubMed, and 27,049 scientific abstracts from Web of Science based on a list of 20 core journals provided by the Institute of Botany, constructing an experimental dataset of over 50,000 articles for knowledge extraction experiments.

[2]Chemical Entities of Biological Interest (ChEBI) is a freely available ontology of biochemical entities focusing on small molecular compounds.

4.2 Design of the Knowledge Extraction Framework

To effectively identify knowledge unit instances and their relationships, we designed the knowledge extraction framework shown in [Figure 3: see original paper]. The framework comprises: (1) Input data sources, including scientific literature for extraction and domain resources (plant diversity ontologies, NCBI Taxonomy, etc.). (2) Extraction tools and methods, employing various natural

language processing tools (Stanford Parser, Berkeley Parser, etc.) for part-of-speech tagging, syntactic dependency analysis, and semantic parsing. (3) Entity and relation extraction, implemented as an iterative cross-process where newly recognized named entities are added to user dictionaries for subsequent recognition rounds, and relation extraction results help discover new entities for further relation discovery. (4) Storage of extraction results, using both RDF storage and database storage depending on result types.

4.3 Knowledge Unit Instance and Relationship Extraction

To achieve rapid and accurate extraction of knowledge unit instances and relationships, we employed dictionary-based, rule-based, and syntactic analysis methods. Dictionary-based entity annotation forms the foundation of all knowledge extraction research. Our study primarily relied on dictionaries to extract species names, geographical locations, some chemical elements and compounds, and domain subject terms. This process aligns with existing research and requires no further elaboration. This section focuses on rule-based instance extraction and new instance identification methods.

4.3.1 Knowledge Unit Instance Annotation and Extraction To identify knowledge unit instances beyond dictionary coverage, we designed a dictionary-based approach combining rules and statistical methods. The process includes:

Rule-based knowledge unit identification. Despite varied manifestations of knowledge unit instances in scientific literature, our analysis revealed common patterns in word formation, part-of-speech, and combination methods (e.g., for persons, institutions, numerical information, equipment). Such instances can be effectively recognized through manually crafted rules. We explored the following general process for rapid rule construction: 1) Collect sample instances of a specific category and perform tokenization, sentence segmentation, and part-of-speech tagging. 2) For simple instances like years, dates, experimental data, and descriptive values, construct patterns using morphological rules. 3) Remove function words (prepositions, adverbs) from tokenization results and identify special proprietary vocabulary (category indicators) from frequency analysis. 4) For instances with category indicators, represent each instance as a pattern of part-of-speech and word form, preserving original strings for feature words and non-noun terms. As shown in [Figure 4: see original paper], Token represents segmentation, Token.orth represents orthography, and Token.category represents part-of-speech. Sample patterns are classified first by feature word position (head, middle, tail), then by word form combinations (e.g., strings without prepositions/possessives, with prepositions, with possessives, or both). This yields effective pattern combinations. For universities, learned patterns include: "of NN/NNS" *(NN/NNS denotes capitalized nouns,* indicates multiple NN/NNS),"NN/NNS* "," (NN/NNS)('s)NN/NNS* ", and"NN/NNS* NN/NNS* ". These patterns, converted to finite state machines, enable instance recognition. 5) For instances without category indicators, collect sample sentences containing research element instances and manually annotate them as training

data. After tokenization, part-of-speech tagging, and parsing, extract linguistic features. Identify n context words surrounding research element instances (n is adjustable; following Jiang et al. [19], we set n=4), calculate word frequencies, and select the top three most frequent neighbors. If these appear in over 50% of entries, they are considered semantic pre-context or post-context indicators. Retrieve synonym sets Synset[train] from WordNet for these indicators. For candidate instances, extract their n context words, obtain Synset[test] from WordNet, and calculate similarity between Synset[train] and Synset[test] using formulas (1) and (2). Using maximum similarity between synonym sets rather than direct word similarity accounts for lexical variation, morphology, and spelling differences, reducing similarity underestimation.

Dictionary similarity-based knowledge unit instance identification. While dictionary-based methods cannot identify new instances, dictionaries provide crucial support. Based on word form, part-of-speech, frequency, and syntactic features, we select candidate new knowledge unit instances and calculate edit distance to dictionary instances to identify unknown words.

Syntactic analysis-based knowledge unit instance identification. Instance and relation extraction are iterative processes where relation extraction results help discover new instances. For candidate instances unrecognizable by rules or dictionary similarity, syntactic dependency and grammatical relationships (parallel components) from parsing can be combined with statistical analysis to determine instance semantic types. Specifically, parsers represent sentences as hierarchical syntax trees. Using "Bell, based in Los Angeles, makes and distributes electronic, computer and building products" as an example, its syntax tree is shown in [Figure 5: see original paper] using Penn Treebank [20] tags compatible with most part-of-speech systems. Beyond syntax trees, parsers provide dependency analysis results shown in [Figure 6: see original paper], where parentheses contain instances and keywords like nsubj and partmod indicate specific dependency relations. Syntactic analysis clearly reveals complex noun phrases (NP modules) and internal dependency features, such as conj_{and}(electronic-11, computer-13) indicating coordination. If one entity' s semantic category is known, the other' s can be inferred, enabling instance type annotation.

4.3.2 Relationship Extraction Knowledge unit relationships include co-occurrence, appositive, coordination, factual, and semantic hierarchical relations. Co-occurrence is simplest: two knowledge unit instances co-occurring within a specified window (full text, abstract, sentence) indicates a relationship. As our texts are journal abstracts, we use sentences as co-occurrence windows. This straightforward approach requires no detailed explanation. We focus instead on syntax-based identification of grammatical, factual, and semantic hierarchical relationships.

Appositive and coordination grammatical relation extraction. As described in "Syntactic analysis-based knowledge unit instance identification," new instance recognition leverages appositive and coordination relations. After confirming

two instances' types, coordination relations (and, or) from syntactic parsing establish appositive and coordination links.

Factual relation identification. Factual relations refer to subject-predicate-object structures <S, P, O> that support subsequent reasoning. The identification process involves: 1) Input tokenized, sentence-segmented texts with identified knowledge unit instances. Process each sentence iteratively, creating empty relation triple lists for verbal and non-verbal predicates. 2) Check if each sentence contains multiple knowledge unit instances from [Figure 2: see original paper]. If not, proceed to the next sentence; if two or more exist, continue. 3) Using parser results, construct minimal simple sentences (containing only one subject-verb-object structure without subordinate clauses) from the syntax tree bottom-up, creating simple sentence groups. 4) For each simple sentence, check if it contains defined knowledge unit instances. If not, proceed to the next sentence; if yes, continue. 5) Extract subject-predicate-object relations using dependency parsing to build (subject phrase, predicate verb, object phrase) triples. 6) Analyze these triples to verify at least one knowledge unit instance exists in both subject and object phrases. If instances are split across both, proceed to step 7; if all reside in the same phrase, jump to step 8. 7) If each phrase contains only one instance without semantic shift, build the relation triple and add it to the verb relation list. For semantic shifts, decide whether to retain the triple based on shifted semantics. If multiple instances exist in a phrase, process them combinatorially while addressing coordination-induced ambiguity. 8) Analyze research element instances in relevant phrases using type annotations to determine semantic relationships. 9) Output the verb relation triple list.

Semantic hierarchical relation discovery. These relations primarily involve possessives, fixed patterns, and common expressions (such as, for example, as well as). We extended Hearst patterns [21] with over 20 manually constructed rules to identify hierarchical relations.

4.4 Application of Knowledge Extraction Results

Applying the above methods to over 50,000 literature titles and abstracts yielded 273,668 knowledge unit instances. The distribution of major extraction types is shown in (displaying only types with >100 instances). In addition to these instances, we extracted 133,922 SPO grammatical relations and 35,903 appositive relations. [Figure 7: see original paper] displays partial SPO extraction results.

Leveraging these results and third-party resources, we constructed a semantic retrieval demonstration platform for plant species diversity, providing domain knowledge revelation, semantic annotation, and ontology navigation to validate our approach' s utility and effectiveness. [Figure 8: see original paper] and [Figure 9: see original paper] showcase application demonstrations.

Compared with general biological knowledge extraction, plant species diversity involves more complex knowledge unit types and relationships (e.g., ecological environments, species characteristics, influencing factors). Therefore, frame-

work design must consider more knowledge units from an application perspective, requiring domain-independent extraction methods adapted to diverse instance types.

Conclusion

Based on analysis of biodiversity information extraction research and project requirements for the "Biodiversity Domain Ontology and Semantic Organization Application Demonstration Platform," this study designed a plant species diversity semantic knowledge extraction framework and explored corresponding extraction methods. Focusing on practical application, we emphasized engineering-oriented knowledge organization frameworks and recognition methods, making dictionaries and manually crafted rules essential components. Consequently, inherent limitations of dictionaries and rules somewhat restrict recognition completeness and accuracy, highlighting the need for continued research on refined recognition of various knowledge unit types.

References

[1] Thessen A E, Cui H, Mozzherin D. Applications of Natural Language Processing in Biodiversity Science [J]. *Advances in Bioinformatics*, 2012. DOI: 10.1155/2012/391574.
[2] Naderi N, Kappler T, Baker C J, et al. OrganismTagger: Detection, Normalization and Grounding of Organism Entities in Biomedical Documents [J]. *Bioinformatics*, 2011, 27(19): 2721-2729.
[3] Species [EB/OL]. [2016-04-12]. http://en.wikipedia.org/wiki/Species.
[4] Gerner M, Nenadic G, Bergman C M. LINNAEUS: A Species Name Identification System for Biomedical Literature [J]. *BMC Bioinformatics*, 2010. DOI: 10.1186/1471-2105-11-85.
[5] The NCBI Taxonomy Homepage [EB/OL]. [2016-04-12]. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonom
[6] Page R D M. BioNames: Linking Taxonomy, Texts, and Trees [OL]. http://dx.doi.org/10.7717/peerj.190.
[7] Species 2000 [EB/OL]. [2016-04-12]. http://www.catalogueoflife.org/annual-checklist/2014/.
[8] Akella L M, Norton C N, Miller H. NetiNeti: Discovery of Scientific Names from Text Using Machine Learning Methods [J]. *BMC Bioinformatics*, 2012. DOI: 10.1186/1471-2105-13-211.
[9] The OrganismTagger System [EB/OL]. [2016-04-12]. http://www.semanticsoftware.info/organism-tagger.
[10] Koning D, Sarlar I N, Moritz T. TaxonGrab: Extracting Taxonomic Names from Text [J]. *Biodiversity Informatics*, 2005, 2: 79-82.
[11] Taylor A. Extracting Knowledge from Biological Descriptions [C]//Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases. 1995: 114-119.
[12] Tang X, Heidorn P B. Using Automatically Extracted Information in Species Page Retrieval [C]//Proceedings of TDWG 2007. 2007.

[13] Cui H. CharaParser for Fine-grained Semantic Annotation of Organism Morphological Descriptions [J]. *Journal of the Society for Information Science and Technology*, 2012, 63(4): 738-754.
[14] Duan Yufeng, Huang Sisi. Information Extraction from Chinese Plant Species Diversity Description Text [J]. *New Technology of Library and Information Service*, 2016(1): 87-96.
[15] Li C, Liakata M, Rebholz-Schuhmann D. Biological Network Extraction from Scientific Literature: State of the Art and Challenges [J]. *Briefings in Bioinformatics*, 2013. DOI: 10.1093/bib/bbt006.
[16] Skusa A, Rüegg A, Köhler J. Extraction of Biological Interaction Networks from Scientific Literature [J]. *Briefings in Bioinformatics*, 2005, 6(3): 263-276.
[17] Bai Guangzu, He Yuanbiao, Ma Jianxia, et al. Application of Machine Learning with Limited Corpus to Identify Structure of Scientific Abstracts Automatically [J]. *New Technology of Library and Information Service*, 2014(7-8): 34-40.
[18] Xu Zheping, Cui Jinzhong, Qin Haining, et al. On the Architecture of Biodiversity e-Science Infrastructure in China [J]. *Biodiversity Science*, 2010, 18(5): 480-488.
[19] Jiang W, Guan Y, Wang X L. Improving Feature Extraction in Named Entity Recognition Based on Maximum Entropy Model [C]//Proceedings of the 5th International Conference on Machine Learning and Cybernetics. 2006: 2630-2635.
[20] De Marneffe M-C, Manning C D. Stanford Typed Dependencies Manual [OL]. http://nlp.stanford.edu/software/dependencies_{manual}.pdf.
[21] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora [C]//Proceedings of the 14th International Conference on Computational Linguistics, 1992.

Author Contribution Statement

Liu Jianhua: Proposed the overall framework, designed the semantic knowledge extraction framework, participated in implementation and development, wrote the main content, and proofread/revised the final version.
Wang Ying: Participated in framework design, corpus preparation, storage structure design, and data processing.
Zhang Zhixiong: Participated in framework design and provided revision suggestions.
Li Chuanxi: Responsible for implementation and development of extraction functions and provided development documentation.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is stored by the authors and available upon request at liujh@mail.las.ac.cn.

[1] Liu Jianhua, Wang Ying, Li Chuanxi. Top40000 SPO Extraction Results.xls.

Received: April 14, 2016
Revised: August 12, 2016

Note: Figure translations are in progress. See original paper for figures.