Data & Knowledge Engineering 100 (2015) 34–53

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Supporting knowledge discovery for biodiversity

Manuel Vilares *, Milagros Fernández, Adrián Blanco

Department of Computer Science, University of Vigo, Campus As Lagoas s/n, 32004 Ourense, Spain

ARTICLE INFO

Article history:
Received 1 September 2014
Accepted 26 August 2015
Available online 5 September 2015

Keywords:
Knowledge discovery
Natural language processing
Text mining

ABSTRACT

A proposal for text mining as a support for knowledge discovery on biological descriptions is introduced. Our aim is both to sustain the curation of databases and to offer an alternative representation frame for accessing information in the biodiversity domain. We work on raw texts with minimum human intervention, applying natural language processing to integrate linguistic and domain knowledge in a mathematical model that makes it possible to capture concepts and relationships between them in a computable form, using conceptual graphs. This provides a reasoning basis for determining semantic disjointedness or subsumption, as well as sub and super-concept relationships.

1. Introduction

Biodiversity description provides basic understanding for decision-making about conservation and sustainable use, affecting a wide range of sectors of both human and economic importance, such as the chemical and agri-food industries. This supports the interest of *taxonomy*, the science of describing, naming and classifying living organisms in an ordered system of *taxa*, largely considered to be unfashionable. So, it is often supposed that DNA barcoding is the ultimate solution to taxa identification, when in fact the arguments in its favor are illusory even for its proponents [24]. People also assume that identifying species is a straightforward and low cost task, but that is far from being the case. We have only to realize that about 2 million species have been documented so far, which means that 80–90% of life is still to be discovered [63]. Furthermore, taxonomists not only delimitate taxa annotating descriptions for new species but also continually refine and review existing ones. Given that the data are distributed across thousands of journals and are provided by different researchers possibly using different vocabularies and methodologies, pursuing particular goals and working under varying spatio-temporal frames [13], taxonomy becomes a complex task of knowledge management. As a result, there is a pressing need for capturing all this information in a way that is semantic, extensible and broadly accessible, which naturally leads us to ontologies [55]. Unfortunately, their generation is too labor-intensive and time-consuming to ever be fully automated, the process relying on qualified experts also known as *curators*.

With respect to access to information, not much has changed since the days of Linnaeus [18], who proposed the use of decision trees to identify taxa. Baptized as *keys*, their generation is a task reserved to curators and, while can be integrated in ontologies using "is-a" links, they have their weaknesses [59]. So, some characteristics may have been omitted in the key due to error or absence at the moment the description was made, such as fruit properties, making them difficult to use. Also, identification can follow an unsuccessful path through the key, either due to the atypical nature of the specimen or to an error in determining whether it meets a decision criterion. This requires a return to the correct path, which is not a trivial task, especially for non-expert users.

We can then conclude that textual descriptions are not only of interest for database curation, but also for identifying species regardless of the user's expertise. Thus, *knowledge discovery* (KD) facilities are increasingly necessary to support manual work, which justifies the interest in *text mining* (TM) techniques to perform *knowledge extraction* (KE) tasks [15].

* Corresponding author. Tel.: +34 988 387280; fax: +34 988 387001.
E-mail addresses: vilares@uvigo.es (M. Vilares), mfgavilanes@uvigo.es (M. Fernández), adbgonzalez@uvigo.es (A. Blanco).

http://dx.doi.org/10.1016/j.datak.2015.08.002
0169-023X/

2. The state-of-the-art

Roughly speaking, tm refers to the process of deriving new knowledge from text, which is often interpreted as comprising three major tasks, namely *information retrieval* (IR), *information extraction* and *data mining*. We can distinguish two approaches: *co-occurrence* and *natural language processing* (NLP) based TM.

2.1. Co-occurrence-based text-mining

Associations between terms are inferred on the assumption that when present in the same sentence or abstract they are related, following a semantic model known as *bag-of-words* (BOW) [27]. The meaning of a text is represented by the multiset of its terms assuming full independence between them. Algebraic [50] and probabilistic [37] approaches are mainly used but, since little attention is paid to the linguistic structure, the type of association is neither identified nor negation dealt with, and thus non-meaningful relationships can arise. To minimize the latter, authors apply weighting criteria to rank the associations, such as *term frequency* (TF), *inverse document frequency* (IDF) and document length [49]. All the above limits the potential interest of this approach for exploratory tm tasks, it now being used more as a baseline method against which others are compared [66].

2.2. NLP-based text-mining

Co-occurrence provides recall, but we need access to a wealth of background knowledge in order to improve precision [31]. This places us within the context of NLP techniques, where syntactic and semantic analyses are combined with morphological and lexical variation through *part-of-speech tagging* (POST), to reveal relationships.

2.2.1. Syntactic modeling

We distinguish three models on the basis of the strategy applied to represent the meaning: *semantic*, *constraint-based logical* and *dependency grammars* (DG). The former [6] fills semantic templates according to sentence patterns. Most proposals [53] rely on *context-free grammars*, including *regular* ones [39]. The lack of contextual sensitivity favors non-determinism, often reduced by the consideration of restrictive sublanguages [20] or domain-specific heuristics [52], which do not go to the heart of the problem and impose the use of specialized grammars. This justifies the interest in formalisms with richly structured lexicons, such as *head-driven phrase structure grammars* [11], although their applicability is questionable when the elements involved in the relevant constructions are not definable in strongly configurational terms [36]. Alternatively, *mildly context-sensitive grammars* (MCSG) have acquired popularity in the sphere of NLP [43] due not only to their lexical sensitivity [51], but also to their capacity to deal with certain cross- and long-distance dependencies in polynomial time and space through the treatment of non-determinism in dynamic programming [14]. This makes it possible to save all parses, postponing the resolution of ambiguities to a semantic stage.

Logical approaches look for the expressiveness of *first-order logic* (FOL) through rules associating predicates and semantic constraints by unification, providing parsing as deduction. The most popular one [42,59] refers to *definite clause grammars* [45], which pose problems of maintenance due to the fixed arity in predicates, meaning that if we wish to extend a grammar each rule must be changed.

Both semantic and constraint-based logical grammars serve as a kernel for *phrase structure parsers*, which break sentences into constituents and can lead to complex structures that neither adapt well to languages with free term order [10], nor look for relationships close to semantic interpretation [22]. In contrast, *dependency parsing* captures the relations between a term and its dependents, simplifying the description and extending [21] the use of DG [60]. However, polynomial time is only achieved in certain cases [25], which suggests that tm should combine information from both dependencies and constituents, looking for a trade-off between syntactic information and ease of phrase extraction. Here we can take advantage of the lexicalized *tree adjoining grammars* (TAG) [33], a type of MCSG for which the derivation controller can be interpreted as a dependency graph [7], allowing the modeling of a dependency parser from rich constituency information. To give this approach a practical sense it is necessary to reduce the combinatorial explosion of trees associated to lexicalization and extended domain of locality, which can be solved by means of tree factorization [14].

2.2.2. Semantic modeling

We seek to support searching and reasoning facilities, but at the same time express content in a form that is logically precise, humanly readable and computationally tractable [57]. This takes us away from formalisms such as *region algebras* [8], which require structured texts [40], and leads us to focus on the so called knowledge-based ones: *description logics* (DLS) and *network-based systems*. The former [2] use a variant of fol, in which reasoning amounts to verifying logical consequence, which provides a decidable and declarative basis for KD. In network-based proposals, knowledge is represented by means of graph-like structures, and reasoning is accomplished by procedures that manipulate them. We here include *semantic networks* [47] and *frames* [38], both of which suffer from the absence of a well defined semantics that translates into a lack of declarative power [5], including difficulty in handling negation. More recently, *conceptual graphs* (CG) [56] have the expressing power of fol. This justifies the consideration of decidable fragments such as the *simple conceptual graphs* (SG), which correspond to existentially quantified conjunctions of atoms. Reasoning is then introduced on the basis of a graph morphism called *projection* [3], which proves to be both sound and complete with regard to deduction.

The graph structure of CG seems to provide a greater expressiveness than the tree one of most DL [16], with two substantial differences between both formalisms. The former refers to the incorporation of both a terminological and an assertional language