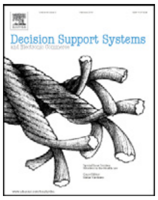
Decision Support Systems 62 (2014) 78–93

Contents lists available at ScienceDirect

Decision Support Systems

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



A semantic approach for extracting domain taxonomies from text

Kevin Meijer, Flavius Frasincar *, Frederik Hogenboom

Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 12 December 2012
Received in revised form 17 March 2014
Accepted 19 March 2014
Available online 27 March 2014

Keywords:
Taxonomy learning
Word sense disambiguation
Term extraction
Subsumption method
Semantic taxonomy evaluation

ABSTRACT

In this paper we present a framework for the automatic building of a domain taxonomy from text corpora, called Automatic Taxonomy Construction from Text (ATCT). This framework comprises four steps. First, terms are extracted from a corpus of documents. From these extracted terms the ones that are most relevant for a specific domain are selected using a filtering approach in the second step. Third, the selected terms are disambiguated by means of a word sense disambiguation technique and concepts are generated. In the final step, the broader–narrower relations between concepts are determined using a subsumption technique that makes use of concept co-occurrences in a text. For evaluation, we assess the performance of the ATCT framework using the semantic precision, semantic recall, and the taxonomic *F*-measure that take into account the concept semantics. The proposed framework is evaluated in the field of economics and management as well as the medical domain.

1. Introduction

In a world where the amount of digital data grows over more than 50% per year, any means to structure this data becomes increasingly relevant [19]. Knowledge management and decision making tasks more and more rely on such unstructured data and its derived, structured knowledge. One way to deal with the growing amount of data is by using taxonomies. A taxonomy is a concept hierarchy in which the broader–narrower relations between different concepts are stored. Taxonomies have proven useful for information search, classification, navigation, etc. [2], and hence can be exploited in decision support systems.

Manually creating a taxonomy, however, remains a difficult and time consuming process. In order to be able to construct high quality taxonomies, a massive amount of knowledge is required [5,13]. Even if the required knowledge is available, it remains a tedious task to organize a high number of concepts in a proper manner. Therefore it is interesting to find ways to automatically build taxonomies [40]. Based on the availability of large text corpora one can investigate the construction of taxonomies from text, using techniques stemming from the closely-related field of terminology engineering [1,10,11,24]. Such automatic taxonomy construction can greatly support the knowledge acquisition phase during the development of a knowledge-intensive decision support system. As the knowledge acquisition is fully automatic, it seamlessly provides up-to-date knowledge in a decision process, which can be of use in real-time business in a wide variety of tasks. For instance, taxonomies can support query formulation targeting at for instance finding articles on a certain theme [3], or can support recommendation systems [39]. Moreover, (automatically built) taxonomies can be employed in faceted search applications [43], or they can be used for summarizing information from different text-based data sources [7]. Last, a common application of taxonomies is in the filtering, enriching, or improving the quality of the data used in support systems [12,25].

To automatically create a taxonomy from text corpora, first, terms need to be extracted. These extracted terms form the lexical representations of the concepts of the taxonomy that is to be built. After the concept lexical representations are determined, the concepts need to be stored in a concept hierarchy to form a taxonomy, which requires the use of clustering techniques. An intermediate step called word sense disambiguation (WSD) may also be applied to ambiguous simple terms. WSD is the process of deriving the sense in which terms are used in a text. For example, the term 'return' may refer to a tennis stroke, but also to a return of money arising from economic transactions. By applying WSD, the taxonomy terms thus have associated a meaning which removes their possible ambiguity. This disambiguation allows for improved concept definition.

A common way of evaluating automatically built taxonomies is by applying a golden standard evaluation [21,9], in which a constructed taxonomy is compared to a benchmark taxonomy. In the past such evaluation, however, only has taken place on a lexical level. As the terms in the benchmark taxonomy are ambiguous, evaluation is limited to comparing the lexical representations of taxonomy concepts. Because of these representations, it might occur that taxonomy concepts that are having the same lexical representations but that are semantically different, are considered to be the same. To prevent this situation, one can apply a semantic comparison of taxonomies. For this purpose, the

* Corresponding author. Tel.: +31 10 408 1340; fax: +31 10 408 9162.
E-mail addresses: kmeijer@hotmail.com (K. Meijer), frasincar@ese.eur.nl (F. Frasincar), fhogenboom@ese.eur.nl (F. Hogenboom).

http://dx.doi.org/10.1016/j.dss.2014.03.006
0167-9236/© 2014 Elsevier B.V. All rights reserved.

concepts of the benchmark taxonomy first need to be disambiguated. In our current endeavors, we present an approach that enables the taxonomy evaluation on a semantic level. In order to be able to apply a semantic evaluation, on both the constructed taxonomy and the benchmark taxonomy, WSD is applied. The main focus of this work is on the use of WSD in the process of automatic taxonomy construction.

In this paper, we present a framework using a semantic approach for the automatic construction of domain taxonomies, called Automatic Taxonomy Construction from Text (ATCT). In the ATCT framework WSD is incorporated. The text corpora that are used to extract terms are the text corpus of RePub[1] and the text corpus of RePEc.[2] RePub is a repository of documents from the Erasmus University Rotterdam, the Netherlands. It contains documents from different domains such as economics, health, law, and psychology. RePEc is an online database which solely contains economic articles collected by volunteers from 76 countries. We have selected both corpora because they are tagged specifically for two domains of interest, i.e., economics and management, and health and medicine.

Two taxonomies are constructed, one for the domain of economics and management, and the other one for the domain of health and medicine. The taxonomy that is constructed for the domain of economics and management uses a total of 25,000 documents from RePub and RePEc. The medicine and health taxonomy uses a total of 10,000 documents from RePub only.

Furthermore, we introduce a new method for disambiguating taxonomy concepts. The application of this method allows for the semantic evaluation of the built taxonomy. The taxonomy for economics and management is semantically evaluated using the STW Thesaurus for Economics and Business Economics[3] as the benchmark taxonomy. The taxonomy for medicine and health on the other hand is evaluated using the MeSH taxonomy,[4] which is a large ontology used for arranging medical subject headings.

The contributions of this paper are six-fold. First, we provide a semantic approach for taxonomy construction from text. Second, we define new evaluation measures, i.e., the semantic precision and semantic recall. Third, the framework as presented in the paper makes use of WSD for both the text corpus as well as the reference ontology (used for evaluation) in order to better define the meaning of concepts. Fourth, we investigate taxonomy construction from text corpora for the field of economics and management, a domain which has not been previously considered for this task in the literature. Fifth, we present a detailed evaluation of the different steps used in our taxonomy construction framework. Last, we refine an existing subsumption method [36] using concept semantics.

The rest of the paper is structured as follows. First, related work in the area of automatic taxonomy construction from text corpora is reviewed in Section 2. Then, the ATCT framework and its implementation are introduced in Sections 3 and 4. Subsequently, the taxonomies built using our ATCT implementation are evaluated in Section 5. Last, we provide a summary of our research, as well as future work directions in the field of automatic taxonomy construction from text in Section 6.

2. Related work

In this section we discuss the current body of literature in the field of automatic taxonomy construction from text. A vast amount of research has been done in this area, and existing works differ in various ways. In general, three different aspects of taxonomy extraction can be distinguished, which are addressed in this section. For each of these aspects we infer the main approaches. First, various methods that have been applied to extract the terms used in taxonomies are described. Then, a review of methods to construct the broader–narrower relation between concepts is presented. Last, previous work concerning the evaluation of the built taxonomies is given. Also, we elaborate on word sense disambiguation techniques that can be applied in automatic domain taxonomy construction processes, and last, we summarize the section with a general discussion on existing extraction methods with respect to our proposed methodology.

2.1. Term extraction

Several methods are available to extract terms from a set of documents. These methods can be broadly categorized into three different approaches: linguistic approaches, statistical approaches, and hybrid approaches.

Linguistic methods use natural language processing (NLP) for term extraction. A linguistic method is part-of-speech tagging [42]. A part-of-speech (POS) tagger labels the part-of-speech (e.g., adjective, noun, verb, etc.) of terms appearing in a text. Another technique is morphological analysis. This technique is used to derive a term's form, e.g., whether a term is used in singular or plural form, the term's inflection, etc. One can also extract terms by using lexico-syntactic patterns, which analyze relations between terms to possibly retrieve new terms [17]. An important feature of linguistic techniques is their ability to define the grammatical functions of terms in sentences. When extracting terms for a certain domain, they however do not consider the relevance of a term for that domain.

Cimiano et al. [6] propose a novel linguistic approach that specifically focuses on verbs. The authors assume that verbs limit the semantic content of their arguments, and hence can be exploited for building conceptual hierarchies by using the inclusion relations between the extensions of the verbs' selectional restrictions. The discussed method relies solely on generic NLP tools for determining the part-of-speech, and hence can be classified as a linguistic method.

Differently than the linguistic approaches, statistical methods do not use the linguistic characteristics of terms, but rely solely on statistical measures to extract terms. These statistical methods are applied to acquire the relevance of a term for a domain. One popular statistical method is the term frequency - inverse document frequency (TF-IDF) [34] measure. This method uses the frequency of a term in a domain corpus document (the term frequency) and the inverse number of corpus documents in which the term appears (the inverse document frequency). The higher the term frequency is in comparison with the document frequency, the more relevant a term is according to the TF-IDF measure. It might occur that a relevant term appears often in corpus documents and thus might not be selected as a relevant term. To prevent such a situation a non-stopping word list can be used [15], on which terms are listed that should never be filtered out.

The authors of [45] provide an example of frequency-based taxonomy extraction for mining characteristic phrases (i.e., sequential patterns) that describe documents. In their extraction phase, meaningless sentences are removed, based on the amount of occurrences within the same paragraph. Another example of a statistical approach to term extraction that does not exploit linguistic characteristics is presented in [26]. Maedche and Volz extract terms from text using several statistical and data mining-based algorithms, mainly based on term frequencies. The outputs of these algorithms are subsequently used for creating concepts and their lexical representations, which can be used in following steps for deriving concept hierarchies. Alternatively, Google page counts can be used [27]. These page counts serve as a substitute for term frequencies, and appear to work well when used for calculating term dependencies, and subsequently adjacencies (resulting in a taxonomy).

Hybrid extraction techniques combine linguistic techniques and statistical measures. An example of a hybrid method is the term filtering method presented in [38]. First, linguistic processing takes place, after which terms are filtered on multiple criteria, e.g., domain pertinence,

[1] Available at http://repub.eur.nl/.

[2] Available at http://repec.org/.

[3] Available at http://zbw.eu/stw/.

[4] Available at http://onto.eva.mpg.de/obo/mesh.owl.