Data & Knowledge Engineering 104 (2016) 1–14

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

An automatic method for reporting the quality of thesauri

Javier Lacasta[a,*], Gilles Falquet[b], F. Javier Zarazaga-Soria[a], Javier Nogueras-Iso[a]

[a]Computer Science and Systems Engineering Dept., Universidad de Zaragoza, Spain
[b]Centre Universitaire d'Informatique, Université de Genève, Switzerland

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 10 March 2016
Accepted 9 May 2016
Available online 16 May 2016

Keywords:
Thesaurus
Digital libraries
Information retrieval
Thesaurus quality
Ontology alignment

ABSTRACT

Thesauri are knowledge models commonly used for information classification and retrieval whose structure is defined by standards such as the ISO 25964. However, when creators do not correctly follow the specifications, they construct models with inadequate concepts or relations that provide a limited usability. This paper describes a process that automatically analyzes the thesaurus properties and relations with respect to ISO 25964 specification, and suggests the correction of potential problems. It performs a lexical and syntactic analysis of the concept labels, and a structural and semantic analyses of the relations. The process has been tested with Urbamet and Gemet thesauri and the results have been analyzed to determine how well the proposed process works.

1. Introduction

In information retrieval (IR) systems, resources are frequently classified using thesauri or other simple knowledge models. The main reasons for this generalized use are their simple structure, the existence of established standards [1] and the integrated support provided by most catalog tools. The use of fully formalized knowledge structures (such as ontologies) is an alternative for these models, but they are more difficult to create and maintain and, in many contexts, there is no a real need of the additional functionality they provide.

The construction of a thesaurus in an area of knowledge requires a careful selection of the concepts and their interrelations in an appropriate general to specific hierarchy [2]. Standards such as ISO 25964 [1] describe the main features the concepts and relations must have. However, the lack of experience in their creation, the time and costs savings, or the over-adaptation to a data collection produce models with heterogeneous concepts and relations [3]. It is relatively common to find thesauri with terminological heterogeneity, overload of specificity, or even lexical issues in concept labels. Additionally, since their relations are too generic, they usually contain unclear hierarchies that are difficult to interpret [3,4]. For example, in the Urbamet thesaurus [5] *car* and *vehicle* are in different branches. Whereas the first concept is considered as a "mean of transport", the second one is considered as a part of the "transportation management". In this context it would be better if *car* were the narrower term of *vehicle* in the "mean of transport" branch. Additionally, in the "transportation management" branch, the *vehicle* concept could be replaced with *vehicle management* or a similar concept.

[*] Corresponding author.
E-mail addresses: jlacasta@unizar.es (J. Lacasta), Gilles.Falquet@unige.ch (G. Falquet), javy@unizar.es (F. Zarazaga-Soria), jnog@unizar.es (J. Nogueras-Iso).

http://dx.doi.org/10.1016/j.datak.2016.05.002
0169-023X/© 2016 Elsevier B.V. All rights reserved.

Many thesauri issues become irrelevant in their original context, because users are accustomed to them. However, they may be a problem for other organizations and casual users that want to reuse these thesauri in their applications, or to use them as a starting point for building more formal knowledge models (e.g., formal ontologies). In these cases of reuse, the quality of thesauri must be evaluated so that the defects can be corrected.

This paper proposes a process that advances in the detection of the syntactic and semantic quality with respect to ISO 25964 specification [1]. It performs a lexical and syntactic analysis of the concept labels, and a structural and semantic analysis of the relations. Lexical and syntactic analysis allow identifying description issues such as the use of acronyms in preferred labels or the detection of concepts that describe too complex ideas. Structural analysis focuses on detecting sets of incompatible relations (such as cycles) and mandatory properties that are not provided. Finally, semantic analysis studies the suitability of the broader/narrower (BT/NT) relations. This semantic analysis is done by looking for compatible relations in WordNet [6] and DOLCE [7] ontologies. The alignment of thesauri with formal models has been used in the last years to increase thesaurus semantics. Since ontologies provide a more precise definition of their concepts and relations, they can be used to replace some items (concepts or relations) of a thesaurus with more precise ones. In this paper, we describe how to use this approach to analyze the quality of thesaurus relations. The proposed process works with both monolingual and multilingual thesauri, but some of the steps used to perform the alignment between thesauri and WordNet/Dolce take advantage of multilingual labels of concepts to improve the matching.

The paper is structured as follows. Section 2 describes the concept of quality in thesauri and the features that are relevant to measure it. Section 3 introduces the proposed analysis method. Section 4 shows the quality analysis of Urbamet and Gemet thesauri. Section 5 reviews other approaches related to thesauri quality detection and compares them with our approach. The paper ends with a discussion about the process, some conclusions, and an outlook on future work.

2. Background in thesaurus quality measures

According to ISO 8402 [8], "the quality" is a measure of excellence or a state of being free from defects, deficiencies and significant variations. It defines the quality as "the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs".

The main sources to identify the quality features of a thesaurus are the existent construction guidelines. They range from practice manuals such as Aitchison et al. [9], to the current international standard ISO 25964 [1]. Kless and Milton [10] aggregate the quality notions in thesaurus literature and they describe a range of abstract measurement constructs, which allow an empirically testable evaluation. It classifies these quality measures according to the different parts of the thesaurus they affect: concepts, terms, structure and documentation parts, and as a whole. These measures include some that can be transformed into rules to be interpreted in an automatic way, such as the number of words defining the number of concepts, or the degree to which there are no redundant terms in the thesauri. However, most of them, such as the degree in which the concepts are in the scope of the thesaurus or the proportion of relevant concepts of the field that are covered, are too general and abstract to be automated. They can be considered more as a general guide of the aspects that are important in a thesaurus than as specific features that need to be reviewed. Somehow, Pinto [11] obtains similar results through a survey with students, young researchers, librarians and experts in information. The result is a set of quantitative measures indicating the perceived importance of the different aspects of thesaurus quality (structural, functional, formal, and external). As in the previous work, most of the measures are guidelines of general aspects to review, but in this case they are focused on the perception of the user (e.g., perceived thesaurus structure or perceived performance). Finally, Mader and Haslhofer [12] focus on analyzing the relevance of technical aspects for quality, such as the lack of definitions or the existence of cycles in the hierarchical relations (it also includes other features, such as the use as Linked Data, which are outside the thesaurus scope). Then, the importance of these features for the quality of a thesaurus is established through a survey with vocabulary managers, term contributors and users of thesauri.

In this paper, to determine the quality of a thesaurus, we have selected the following features:

Property completeness measures:These measures are focused on the identification of lacking properties. We analyze the completeness and uniqueness of preferred labels and completeness of definitions.

Property content measures:Their objective is to locate invalid values inside labels. We focus on detecting non-alphabetic characters, adverbs, initial articles, and acronyms (in preferred labels).

Property context measures:These are focused on identifying anomalies involving several labels. This includes detecting duplicated labels and inconsistencies in the use of uppercase and plurals.

Property complexity measures:They provide a measure of the syntactic complexity of the labels, in terms of the use of prepositions, conjunctions and adjectives.

Relation coherence measures:They indicate if the relations are complete, coherent, and semantically correct. RT analysis focuses on detecting non-informative relations (they link hierarchically related concepts). BT/NT analysis searches for cycles in the model, unlinked concepts and relations that do not associate a superordinate with a subordinate concept. According to ISO 25964, the superordinate must represent a class or whole and subordinate its members or parts.

They are technical features obtained from the rules and recommendations in ISO 25964 standard. These rules include the same thesaurus related features as Mader and Haslhofer [12] and additional ones related to the syntaxis and semantics of the labels, concepts and relations (e.g., the use of adverbs, acronyms, and the meaning of BT/NT relations).