Chapter 11
Semantification of Large Corpora of Technical Documentation

Sebastian Furth
denkbares GmbH, Germany

Joachim Baumeister
denkbares GmbH, Germany & University of Würzburg, Germany

ABSTRACT
The complexity of machines has grown dramatically in the past years. Today, they are built as a complex functional network of mechanics, electronics, and hydraulics. Thus, the technical documentation became a fundamental source for service technicians in their daily work. The technicians need fast and focused access methods to handle the massive volumes of documentation. For this reason, semantic search emerged as the new system paradigm for the presentation of technical documentation. However, the existent large corpora of legacy documentation are usually not semantically prepared. This fact creates an invincible gap between new technological opportunities and the actual data quality at companies. This chapter presents a novel and comprehensive approach for the semantification of large volumes of legacy technical documents. The approach especially tackles the veracity and variety existent in technical documentation and makes explicit use of their typical characteristics. The experiences with the implementation and the learned benefits are discussed in industrial case studies.

INTRODUCTION
In the mechanical engineering domain, the growth of complex products including a number of different technologies can be experienced. In the past, mechanical engines typically could be simply maintained and repaired using mechanical tools. Today’s machines, however, are designed as a combination of mechanical components, electrics, hydraulics, and electronics. Consequently, today’s service technicians need an increased competence for their daily repair and maintenance work. Technical documentation tries to support the service technician with relevant information on the respective machine. However,
as a result of the rising complexity of the products the corpora of technical documentation become extremely large (Volume; typically up to 12,000 pages for a single machine). Terabytes of data are a common dimension, as the documentation usually not only needs to cover different skill levels of users, configurations and aspects of a product but also targets different markets and their respective languages. Hence, documentation exists for each variant of a machine and every relevant market (Variety; different kinds, content structures and languages of documentation). In order to provide efficient customer support fast and effective access to the relevant information becomes a critical success factor.

In recent years, well-established semantic technologies (Antoniou & van Harmelen, 2008; Horrocks, Parsia, Patel-Schneider, & Hendler, 2005; Sauermann, et al., 2006) have been applied to facilitate this task. A typical example is Semantic Search (Guha, McCool, & Miller, 2003) that—in contrast to traditional search technologies—uses ontologies to connect content with semantic information. This connection can then be exploited in order to improve the search results. Semantic Navigation is another example of advanced technologies enabling users to browse multi-modal content by following automatically generated semantic links. However, the incorporation of these technologies in enterprise applications has just started and only a small amount of technical documentation is already semantically enriched. Thus, the service staff usually still has to deal with large quantities of rather loosely organized legacy information. Examples of such information include operation manuals, installation guides, and repair manuals. In order to make the large volumes of legacy information accessible for semantic technologies, connections between the traditional content and semantic information must be created. This is typically done in a process called Ontology Population (Buitelaar & Cimiano, 2008), where a given ontology structure is filled by instances. These instances describe for example the main subject (in terms of ontology concepts) of a document. This kind of ontology population can be regarded as a form of the Entity Integration task in Big Data, and is vaguely related to Subject Indexing (Albrechtshen, 1993; Hutchins, 1978), which in turn can be considered as part of the more general problem of Document Classification.

Creating these instances in a manual step requires an in-depth analysis of the original content by humans, which is usually error-prone, time-consuming, and very cost-intensive for large scale corpora. In recent years, however, Text Analytics approaches have been adapted in order to automate the population of such ontologies.

BACKGROUND

In the field of Information Extraction and Text Analytics, established methods exist for the extraction of semantic information from natural language texts. Most of these methods are based on supervised Machine Learning approaches that require a sufficient amount of training data for decent results. In real-world scenarios such training data is often not available, and the creation with respect to a cost-benefit ratio is not reasonable.

The chapter presents a holistic approach for the automatic semantification of large volumes of technical documentation that does not require training data. The approach is holistic, as it is a complete process that covers all steps necessary for the semantification of existing technical documents. In this chapter, semantification is defined as the identification and annotation of the main subjects for a given document. For instance, the chapter of a repair manual is identified and annotated by ontological subjects representing the main content of the chapter in focus. The typical goal of the presented semantification
Related Content

Improvement of Variant Adaptable LSTM Trained With Metaheuristic Algorithms for Healthcare Analysis
[www.igi-global.com/chapter/improvement-of-variant-adaptable-lstm-trained-with-metaheuristic-algorithms-for-healthcare-analysis/222143?camid=4v1a](www.igi-global.com/chapter/improvement-of-variant-adaptable-lstm-trained-with-metaheuristic-algorithms-for-healthcare-analysis/222143?camid=4v1a)

A Framework to Analyze Enterprise Social Network Data
[www.igi-global.com/chapter/a-framework-to-analyze-enterprise-social-network-data/154556?camid=4v1a](www.igi-global.com/chapter/a-framework-to-analyze-enterprise-social-network-data/154556?camid=4v1a)

Evolution of Subway Network Systems, Subway Accessibility, and Change of Urban Landscape: A Longitudinal Approach to Seoul Metropolitan Area
[www.igi-global.com/chapter/evolution-of-subway-network-systems-subway-accessibility-and-change-of-urban-landscape/197177?camid=4v1a](www.igi-global.com/chapter/evolution-of-subway-network-systems-subway-accessibility-and-change-of-urban-landscape/197177?camid=4v1a)

Unstructured Healthcare Data Archiving and Retrieval Using Hadoop and Drill