Natural Language Processing Techniques for Document Classification in IT Benchmarking

Automated Identification of Domain Specific Terms

Matthias Pfaff[1] and Helmut Krcmar[2]

[1]*fortiss GmbH, An-Institut der Technischen Universität München, Guerickestr. 25, 80805 München, Germany*

[2]*Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany*

Keywords: IT Benchmarking, Natural Language Processing, Heterogeneous Data, Semantic Data Integration, Ontologies.

Abstract: In the domain of IT benchmarking collected data are often stored in natural language text and therefore intrinsically unstructured. To ease data analysis and data evaluations across different types of IT benchmarking approaches a semantic representation of this information is crucial. Thus, the identification of conceptual (semantical) similarities is the first step in the development of an integrative data management in this domain. As an ontology is a specification of such a conceptualization an association of terms, relations between terms and related instances must be developed. Building on previous research we present an approach for an automated term extraction by the use of natural language processing (NLP) techniques. Terms are automatically extracted out of existing IT benchmarking documents leading to a domain specific dictionary. These extracted terms are representative for each document and describe the purpose and content of each file and server as a basis for the ontology development process in the domain of IT benchmarking.

1 INTRODUCTION

Benchmarking as a systematic process for improving organizational performance has gained great popularity worldwide since the 1980s (Camp, 1989). It is based on the insight that analyzing the acting and performance of organizations is a powerful way to transform the own organization. This is done by applying lessons learned for the own organization derived by these observations (Peters, 1994; Camp, 1995). Moreover, this performance measurement (equiv. benchmarking) can help to explain value or cost aspects to stakeholders (Spendolini, 1992). Thus, the analysis and evaluation of such performance measurement approaches is subject of manifold studies (cf. Slevin et al., 1991; Smith and McKeen, 1996; Gacenga et al., 2011).

The research focus of attention is on structuring, standardize and generalize IT service catalogues (cf. Kütz, 2006; Rudolph, 2009; Nissen et al., 2014). Usually, in order to model internally provided (IT) services in a standardized manner. However, since (IT) service catalogues are commonly designed for internal or individual purposes only comparability is difficult to reach, especially across different (IT) organizations. At present, most of research in (IT) benchmarking is focusing on how benchmarking can be done or in how a successfully performed benchmark should be performed (Jakob et al., 2013). In other words, current research on (IT) benchmarking generally focuses on designing service catalogues or designing benchmarks on various kinds of subjects. Due to the nature of the subject, the information collected during a benchmark is generally done by the use of questionnaires. This leads to a variety of different kind of data getting collected withing a single benchmark (such as cost of employee, software licencing costs, quantities of hardware etc.). All of these approaches have one thing in common: A common concept for data management is left out of scope, even though it is strongly recommended (Pfaff and Krcmar, 2014; Wollersheim et al., 2014). Moreover, little work published in IS literature addresses the problem of data integration across different kind of IT benchmarks, yet. So, they omit facts of data quality and data integration.

Today, one difficulty in making data of different types of benchmarking comparable with each other is a result from the lack of a uniform description of any parameter measured. Their relation in between is not formalized too. Following Pfaff and Krcmar (2014) the conceptual level of the different benchmarking approaches needs to be analyzed, to iden-

Pfaff M. and Krcmar H..
Natural Language Processing Techniques for Document Classification in IT Benchmarking - Automated Identification of Domain Specific Terms.
DOI: 10.5220/0005462303600366
In *Proceedings of the 17th International Conference on Enterprise Information Systems* (ICEIS-2015), pages 360-366
ISBN: 978-989-758-096-3

tify first similarities in a logical manner. To do so, already existing service description as well as questionnaires of different benchmarking approaches are used for examination. These data were collected over the last seven years within different benchmarking approaches supervised and evaluated. Encompassing data from strategic and consortial IT benchmarks, reflecting a broad range of numerous small to medium sized enterprises as well as large-scale enterprises.

By the identification of domain specific terms elaborating the specific structural characteristics from different benchmarking approaches, this work addresses the following question: How can the domain specific terms in IT benchmarking be automatically identified out of unstructured data? Subsequently, the results of this work are used to identify the requirements semi-structured and unstructured benchmarking data pose for the use of ontology.

To ensure maximum re-usability and to speed up the document classification process these benchmarking data are analyzed by the use of natural language techniques (NLP). Resulting in a domain specific dictionary as a basis for a domain specific ontology for IT benchmarking, in order to make these kind of data meaningful (Uschold and Gruninger, 2004; Horkoff et al., 2012).

First, an overview of benchmarking in general and data integration challenges in the domain of IT benchmarking in specific is given. Second, the used method and the quality of the previously mentioned approach is described in the following sections. Thus, in this paper the first step in the ontology engineering process is addressed by the use of NLP techniques.

2 RELATED WORK

Today, there exist a broad range of different approaches for structuring service catalogues (cf. Rudolph and Krcmar, 2009). A short overview of these approaches is given by Nissen et al. (2014). Next to IT service catalogues the structure of IT benchmarks follow the abstraction of IT departments proposed by Riempp et al. (2008). Thus, data management in IT benchmarking needs to cover a broad range of different characteristics (e.g. different views on supplier or provider of services, different level of abstraction of a service or various types of cost accounting). Especially where IT-based solutions become more and more used for the data collecting process in the domain for IT benchmarking, such as presented by Ziaie et al. (2012) and structural described by Riempp et al. (2008). Although such benchmarks do have the same object of observation (f.i. same service or same product), no direct semantic information are stored to identify this similarity, which is inhibiting further comprehensive analysis (Pfaff and Krcmar, 2014).

In related fields of research there already do exist several approaches to organise and integrate such kind of semantically identical information. Ontologies which, by definition, convey electronic or ”semantic meaning” are used to structure such kind of unstructured data in the medical sector (cf. Cambria et al., 2011) or in the sector of information management (cf. Riedl et al., 2009; Müller, 2010; Cambria et al., 2011). To address this lack of appropriate data management concept in the domain of IT benchmarking onotlogies are already proposed by Pfaff and Krcmar (2014), following Guarino (1995) and Brewster and O’Hara (2007).

There exist several types of ontology development strategies in academic literature (cf. Wache et al., 2001). A *single ontology* uses a shared vocabulary for describing the semantic information of data. *Multiple ontologies* are based on several independently build ontologies for every source of information. The lack of a shared vocabulary across these ontologies is one major disadvantage. *Hybrid ontologies* use a shared vocabulary with basic terms of the domain related information. But, to our knowledge no ontology exists for IT benchmarking or IT service management.

3 METHODS

Since NLP driven ontology development has become more and more common over the last years, (cf. Lame, 2005; Maynard et al., 2008; Witte et al., 2010; Ray and Chandra, 2012; Karanikolas and Skourlas, 2010; Alatrish et al., 2014) these techniques are used to develop a domain specific ontology for IT benchmarking. Focusing on the first phase of ontology development, such as term extrusion and dictionary development.

3.1 Ontology Development

Ontologies aim to capture static domain knowledge in a generic way and can be used and shared across applications and groups (Chandrasekaran et al., 1999). Thus, one can define an ontology as a shared specification of a conceptualization. Following Noy and McGuinness (2001) and Pinto and Martins (2004) Figure 1 shows the schematic procedure of the ontology creating an process.

First, already existing repositories of information, such as documents, are used to identify and ex-

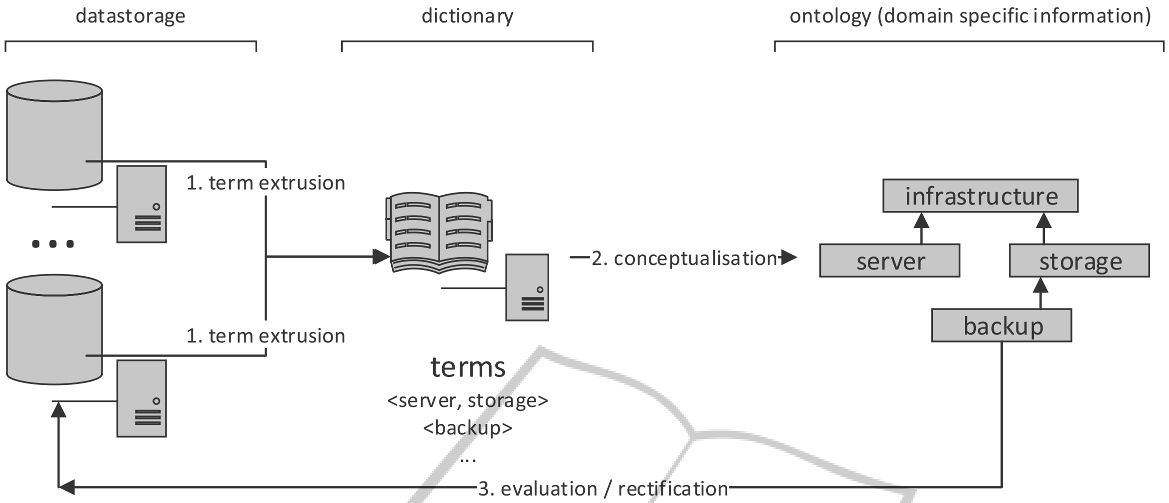


Figure 1: Ontology Engineering steps adapted from Sack (2008).

tract characteristic terms within the specific domain. Second, these terms are conceptualized according to Fernandez-Lopez et al. (1997). In a third step, the conceptualization is evaluated and revised to map the requirements previously identified. Supporting the construction of ontologies and populating them with instantiations of both concepts and relations, commonly referred to as ontology learning.

Next to a manual extraction of terms out of documents there exist several semi-automatic approaches. In general, these are natural language processing (NLP) or machine learning techniques (ML) which speed up the initial process of the ontology engineering.

3.2 Natural Language Processing

Based on already existing documents (i.e. service descriptions and benchmarking results of the last seven years) an automatic extraction of terms is performed. All of the documents stored in various data formats are converted into a new data format, commonly referred to as data stream (raw text). This raw text is the input for the NLP algorithm. Figure 2 illustrates the pipeline architecture for an information extraction system apart from technical details.

The complexity of the NLP analysis can be reduced since all documents are related to topics in the domain of IT benchmarking. It can therefore be assumed that these documents are based on a reduced set of vocabularies. Thus, a dictionary with commonly used terms in this domain supports the NLP process. Using this dictionary a pre classification of the documents can be made according to the initial set of terms. But, as it cannot be assumed that the initial generated dictionary is completely sound, this dictionary has to be iteratively adjusted or extended with the automatically identified terms of the analyzed the documents. As a result a representative set of terms for the domain of IT benchmarking is acquired.

On the pre-processing side of NLP the documents are parsed and transferred into a raw data format which is needed for *tokenization*, *division in sentences*, *lemmatization* and *lexical analysis*. As *tokenization* identifies each single term of a sentence *division in sentences* organizes these terms by grouping them into sentences. The reduction of each term to its basic form is called *lemmatization* (e.g. employees is reduced to employee). In a last step *lexical analysis* aims at the identification of grammatical classes for each term selected in the tokenization process.

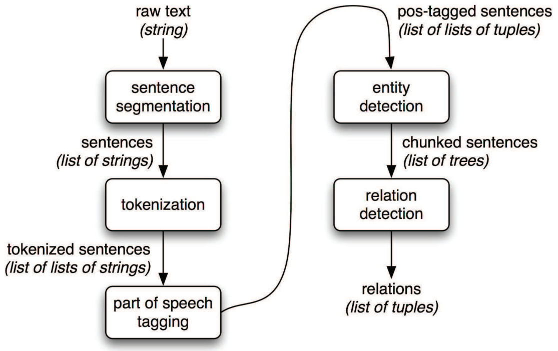


Figure 2: Pipeline Architecture for an Information Extraction System based on Bird et al. (2009).

Following Salton (1989) all words are analyzed and count according to their frequency of use within the existing documents first. The term frequency (t) within on single document (d) is brought into relation of all documents where (t) is used. This is called *inverse document frequency (IDF)*.

$$IDF(t) = \frac{FREQ_{td}}{DOCFREQ_t} \qquad (1)$$

Thus, in a collection of (n) documents the significance (S_{ik}) for one term (t) in document (d) can be described by:

$$S_{ik} = C * \frac{n}{DOCFREQ_t} * FREQ_{td} \qquad (2)$$

Where (C) is known as *Zipf's law* (Zipf, 1949), approximating the rank-frequency relationship where (r) is the rank of a term, (f) is the frequency of occurrence of the term, and (c) is a constant, dependent on the number of terms in a document.

$$C = r * f \qquad (3)$$

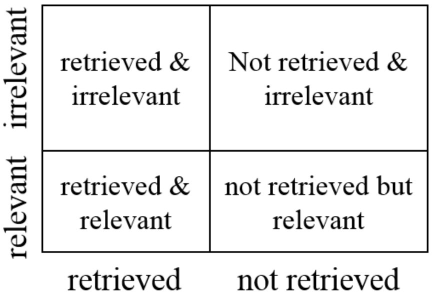
This approach has its weaknesses in small to mid size documents with less different terms. In this case the documents get probably not identified by the most representative term if only the most weighted terms get saved. This will lead to an incomplete list of index terms an therefore inadequate for the building of a base dictionary for IT benchmarking. Consequently, terms of small an mid size documents are parsed last and compared with the dictionary entries created out of larger data sets. In case of new index terms, these terms are included into to dictionary. In case of a document with equivocal results concerning the representative term all terms are stored and associated with this document. This is done in order to prevent incomplete set of dictionary terms as well as incomplete result sets if searched for a specific term and its corresponding documents.

Before measuring the quality and effectiveness of the implemented automated document indexation it is necessary to specify the requirements the implementation has to full fill. In our case these are:

- All relevant information are extracted.
- Less irrelevant information are stored.

Thus, effectiveness reflects the amount of correct identified documents with less false positive results. Moreover, the list of documents identified correct should be nearly complete and the amount of documents not relevant for a specific search term should be small.

The four categories a document can be assigned to is shown in Figure 3. According to the definition of information retrieval systems, an information can be retrieved and be relevant (true positive) or retrieved and irrelevant (false positive). In contrast, the information not received can be irrelevant (false negative) or relevant (true negative).



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

Figure 3: Segmentation of a collection of documents according to four types of classes of belonging (Nohr, 2003).

To measure the effectiveness, two key performance indicator are used, *recall* and *precision* Nohr (2003). *Recall* and *precision* are defined as follows:

$$Recall(r) = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevantdocuments} \qquad (4)$$

$$Precision(p) = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved} \qquad (5)$$

By definition, a high value of *recall* describes a set of documents where all relevant documents are identified, with its drawback, that this set may also contain irrelevant documents. Such high values of *recall* is desired if it is important to identify all documents related to a specific search term. In contrast, a high value of *precision* describes a set of documents with many relevant documents are identified correctly and the amount of irrelevant documents is comparatively low. Thus, a high value of *precision* is desired whenever relevant documents need to be identified only, at the expense of completeness.

4 METHODOLOGY

As already mentioned, it can be assumed, that most of the documents consist of a reduced set of vocabulary, as all of them are related to specific topics out of IT benchmarking. Thus, they describe technical and economic aspects such as IT costs or the number of employees. This constraint allows us to group data objects into subsets based on their relation, i.e. objects with similar information are grouped together.

The reduction to primary words is done by the help of LemmaGen (Juršic et al., 2010; LemmaGen, 2011), a lexical database that contains approximately

23385 natural language terms and about 10655 primary words.

4.1 Prototype

Figure 4 shows the schematic workflow of the implemented prototype. First a set of documents is analyzed according to the previously described NPL methods and transferred into raw data formats. Second, the shared terms of the different documents are identified, building the underlying dictionary of the domain. Therefore LemmaGen (Juršic et al., 2010) and the Stop Word (Savoy, 2014) identifier are used. This shared dictionary is used to identify each single document in a last step (e.g. by name, unit, year and representative tag).

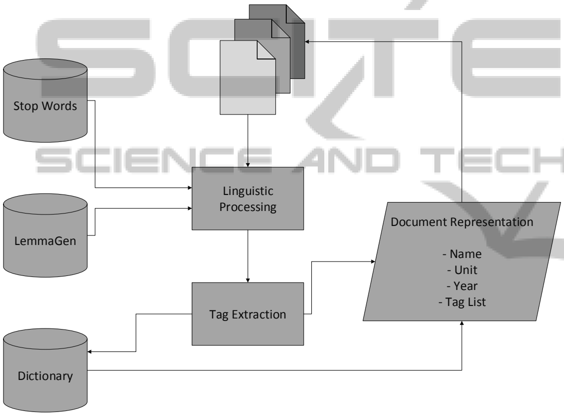


Figure 4: Schematic workflow of the prototype for document indexing.

The implementation of this prototype is done in Java. The documents are read in by the use of the Apache POI API (Foundation, 2014). This is to transform each document into a string-array, split into paragraphs for term identification. At last, each document gets tagged by its most representative term or list of terms.

4.2 Evaluation

According to this schematic workflow the prototype is tested on a set of documents out of different benchmarking approaches, mainly based on *.doc(x), *.xls as well as *.pdf documents, resulting in 1084 unique files. These files were previously categorized by hand, to identify relevant documents with potential terms for ontology building. Moreover, this is done to measure *recall* and *precision*, as the document distribution needs to be known (e.g. documents related to personal costs). This leads to a distribution of documents shown in Table 1.

Table 1: Documents under examination.

Total Number of Documents	1084
Number of relevant Documents	404

At first, the quality of document identification has been tested. Thus, it is evaluated if all relevant documents are found. The results are shown in Table 2.

Table 2: Accuracy of document identification.

Number of relevant documents	404
Number of identified documents	378
Accuracy	93.3%

26 documents could not be identified, as these missed some relevant information needed, such as the name of performance indicator that should be described by this document.

In a next step a subset of manually categorized documents were tested to measure the *precision* and *recall*, while focusing on a high *recall* value. This is due to the fact, that in case of IT benchmarking and especially for the development of an ontology nearly all relevant information/documents should be identified. This means, that false positive identified documents are allowed to occur in the result set. An overview on used search terms is given in Table 3.

Table 3: *Recall* and *precision* for the test data set.

Search term	Recall (%)	Precision (%)
Supported Devices	0.2	1.0
Personnel costs	0.57	0.8
Number of client devices	0.63	1.0
Total cost of IT	0.65	0.92

At last, it is tested whether all units of the indicators are identified correctly. The Result of this test is shown in Table 4. Five units could not be identified because of major typing errors within these documents.

Table 4: Identification of units.

Number of search documents	36
Identified Units	31
Accuracy	0.86%

5 DISCUSSION & FUTURE WORK

This work transfers NP and machine learning techniques into the domain of IT benchmarking, as basis for ontology creation processes in the future. It is its first step towards an ontology in this domain. By automating the term extrusion out of benchmarking documents the development of this ontology is accelerated. This acceleration is even more important on maintaining an ontology. As the initial development of such an ontology is only the first step, extension and maintenance processes are activities which also get supported by the automated term extrusion. This is especially useful if new domain specific terms need to be identified out of new documents, such as service descriptions (e.g. related to topics like cloud computing).

Future work will focus on step two/three, shown in Figure 1. As it is shown, the conceptualization of terms leads, in general, to a cyclically adjustment of the initial developed ontology. As this process needs to be supervised by a domain expert only a semi- automation of this step is possible yet. Nevertheless this semi-automation will be developed. To support the domain expert during this step, the differences between two ontology versions (before and after the automatic term extrusion) will be identified and presented to him. Moreover this kind of versioning helps to comprehend the development process of the whole ontology.

In a last step, already existing output data will be linked to the domain ontology, such as, cost or performance values collected from different companies since the last seven years and persisted in various databases (eg. MySQL or Access DB). Thus, the conceptualization of logical structures in this domain, is used to get access to benchmarking data. Without the need of the development of a unified database schema. Therefore new databases can be linked to already existing ones by the use of an abstraction layer, so called ontology.

REFERENCES

Alatrish, E. S., Tosic, D., and Milenkovic, N. (2014). Building ontologies for different natural languages. *Comput. Sci. Inf. Syst.*, 11(2):623–644.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Brewster, C. and O'Hara, K. (2007). Knowledge representation with ontologies: Present challenges - future possibilities. *International Journal of Human-Computer Studies*, 65(7):563–568.

Cambria, E., Hussain, A., and Eckl, C. (2011). Bridging the gap between structured and unstructured health-care data through semantics and sentics. In *Proceedings of ACM*.

Camp, R. (1989). *Benchmarking: The search for industry best practices that lead to superior performance*. Quality Press, Milwaukee, Wis.

Camp, R. (1995). *Business process benchmarking : finding and implementing best practices*. ASQC Quality Press, Milwaukee, Wis.

Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14(1):20–26.

Fernandez-Lopez, M., Gomez-Perez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium*, pages 33–40.

Foundation, A. S. (2014). Apache poi api. http://poi.apache.org.

Gacenga, F., Cater-Steel, A., Tan, W., and Toleman, M. (2011). It service management: towards a contingency theory of performance measurement. In *International Conference on Information Systems*, pages 1–18.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43(5-6):625–640.

Horkoff, J., Borgida, A., Mylopoulos, J., Barone, D., Jiang, L., Yu, E., and Amyot, D. (2012). *Making Data Meaningful: The Business Intelligence Model and Its Formal Semantics in Description Logics*, volume 7566 of *Lecture Notes in Computer Science*, book section 17, pages 700–717. Springer Berlin Heidelberg.

Jakob, M., Pfaff, M., and Reidt, A. (2013). A literature review of research on it benchmarking. In Krcmar, H., Goswami, S., Schermann, M., Wittges, H., and Wolf, P., editors, *11th Workshop on Information Systems and Service Sciences*, volume 25.

Juršic, M., Mozetic, I., Erjavec, T., and Lavrac, N. (2010). Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Karanikolas, N. N. and Skourlas, C. (2010). A parametric methodology for text classification. *Journal of Information Science*, 36(4):421–442.

Kütz, M. (2006). *IT-Steuerung mit Kennzahlensystemen*. dpunkt.verlag, Heidelberg.

Lame, G. (2005). Using nlp techniques to identify legal ontology components: Concepts and relations. In Benjamins, V., Casanovas, P., Breuker, J., and Gangemi, A., editors, *Law and the Semantic Web*, volume 3369 of *Lecture Notes in Computer Science*, pages 169–184. Springer Berlin Heidelberg.

LemmaGen (2011). LemmaGen, multilingual open source lemmatisation framework. http://lemmatise.ijs.si/Services.

Maynard, D., Li, Y., and Peters, W. (2008). Nlp techniques for term extraction and ontology population. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, pages 107–127, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Müller, M. (2010). *Fusion of Spatial Information Models with Formal Ontologies in the Medical Domain*. Thesis.

Nissen, V., Petsch, M., Jung, D., and Praeg, C.-P. (2014). *Empfehlungen fr eine generelle IT-Service-Katalog-Struktur*, book section 8, pages 133–154. Springer Fachmedien Wiesbaden.

Nohr, H. (2003). *Grundlagen der automatischen Indexierung: ein Lehrbuch*. Logos-Verlag.

Noy, N. F. and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.

Peters, G. (1994). *Benchmarking Customer Service*. Financial Times Management Series. McGraw-Hill, London.

Pfaff, M. and Krcmar, H. (2014). Semantic integration of semi-structured distributed data in the domain of it benchmarking. In *16th International Conference on Enterprise Information Systems (ICEIS)*.

Pinto, H. S. and Martins, J. P. (2004). Ontologies: How can they be built? *Knowledge and Information Systems*, 6(4):441–464.

Ray, S. and Chandra, N. (2012). *Building Domain Ontologies and Automated Text Categorization: a contribution to NLP*. LAP Lambert Academic Publishing.

Riedl, C., May, N., Finzen, J., Stathel, S., Kaufman, V., and Krcmar, H. (2009). An idea ontology for innovation management. *International Journal on Semantic Web and Information Systems*, 5(4):1–18.

Riempp, G., Müller, B., and Ahlemann, F. (2008). Towards a framework to structure and assess strategic IT/IS management. *European Conference on Information Systems*, pages 2484–2495.

Rudolph, S. (2009). *Servicebasierte Planung und Steuerung der IT-Infrastruktur im Mittelstand: Ein Modellansatz zur Struktur der IT-Leistungserbringung in mittelstndischen Unternehmen*. Thesis.

Rudolph, S. and Krcmar, H. (2009). Maturity model for it service catalogues an approach to assess the quality of IT service documentation. pages 759–759.

Sack, D. H. (2008). *Semantic Web*. Hasso-Plattner-Institute, Potsdam.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Savoy, J. (2014). German stopwordlist. http://members.unine.ch/jacques.savoy/clef/germanST.txt.

Slevin, D. P., Stieman, P. A., and Boone, L. W. (1991). Critical success factor analysis for information systems performance measurement and enhancement: A case study in the university environment. *Information & management*, 21(3):161–174.

Smith, H. A. and McKeen, J. D. (1996). Measuring is: how does your organization rate? *ACM SIGMIS Database*, 27(1):18–30.

Spendolini, M. J. (1992). *The benchmarking book*. Amacom New York, NY.

Uschold, M. and Gruninger, M. (2004). Ontologies and semantics for seamless connectivity. *SIGMOD Record*, 33(4).

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In Stuckenschmidt, H., editor, *IJCAI–01 Workshop: Ontologies and Information Sharing*, pages 108–117.

Witte, R., Khamis, N., and Rilling, J. (2010). Flexible ontology population from text: The owlexporter. In *In: Int. Conf. on Language Resources and Evaluation (LREC*.

Wollersheim, J., Pfaff, M., and Krcmar, H. (2014). Information need in cloud service procurement - an exploratory case study. In *E-Commerce and Web Technologies - 15th International Conference, EC-Web 2014, Munich, Germany, September 1-4, 2014. Proceedings*, pages 26–33.

Ziaie, P., Ziller, M., Wollersheim, J., and Krcmar, J. (2012). Introducing a generic concept for an online IT-Benchmarking System. *International Journal of Computer Information Systems and Industrial Management Applications*, 5.

Zipf, G. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley Press.