Understanding the Cloud Computing Ecosystem: Results from a Quantitative Content Analysis

Benedikt Martens
University of Osnabrueck
Accounting and Information Systems
Katharinenstr. 1
49069 Osnabrueck
+49 541 969 4524
benedikt.martens@uni-osnabrueck.de

Jens Poeppelbuss
University of Muenster
European Research Center for Information Systems (ERCIS)
Leonardo-Campus 3
48149 Muenster
+49 251 83 38069
jens.poeppelbuss@ercis.uni-muenster.de

Frank Teuteberg
University of Osnabrueck
Accounting and Information Systems
Katharinenstr. 1
49069 Osnabrueck
+49 541 969 4961
frank.teuteberg@uni-osnabrueck.de

ABSTRACT

An increasing number of companies make use of Cloud Computing services in order to reduce costs and increase flexibility of their IT infrastructure. This has enlivened a debate on the benefits and risks of Cloud Computing, among both practitioners and researchers. This study applies quantitative content analysis to explore the Cloud Computing ecosystem. The analyzed data comprises high quality research articles and practitioner-oriented articles from magazines and web sites. We apply n-grams and the cluster algorithm k-means to analyze the literature. The contribution of this paper is twofold: First, it identifies the key terms and topics that are part of the Cloud Computing ecosystem which we aggregated to a comprehensive model. Second, this paper discloses the sentiments of key topics as reflected in articles from both practice and academia.

Keywords

Cloud Computing, Quantitative Content Analysis, Sentiment Analysis

1. INTRODUCTION

Over the recent years, Cloud Computing has emerged as a new computing paradigm which aims to provide reliable, customized, high-quality and dynamic computing services for end-users [38]. In 2006, Amazon launched their new business division Amazon Web Services (AWS) and provided the basis for this practitioner-driven phenomenon [13]. Cloud Computing utilizes existing technologies like Grid Computing and Virtualization for the delivery of scalable IT services via the internet on a pay-per-use basis [39]. Nevertheless, the technologies employed for Cloud Computing are still in the process of maturing [25,38]. Also, definitions, attributes and characteristics associated with Cloud Computing will continue to evolve and change over time [26].

The three main types of Cloud Computing services are: Software as a Service (SaaS), which refers to application services like Salesforce; Platform as a Service (PaaS), e. g., developer platforms like the Google AppEngine; and finally Infrastructure as a Service (IaaS), which mainly encompasses storage services and computing power services like AWS [25,39].

The concept of Cloud Computing receives increasing attention in both academia and practice [18,23,25]. It attracts researchers and engineers from various backgrounds (e. g., economic vs. technical) who approach the topic from different perspectives (e. g., provider vs. customer). Generally, the overall trend seems to be that of continuously growing interest in Cloud Computing and associated topics like IT Outsourcing, Grid Computing, and Virtualization. This impression was confirmed by the results of a analysis of Google Insights for Search we conducted (cf. Figure 1). To make the data comparable to each other, they are normalized on a scale of 0 and 100. Each point on the graph has been divided by the highest one, or 100.

It becomes obvious that until the third quarter of 2007 the number of search queries regarding the term “Cloud Computing” was on a constant increase. The interest in “Grid Computing” and “IT Outsourcing” slackened until the middle of 2008 and remained more or less steady from then. In contrast, until the beginning of 2010 there was a recognizable upward trend in the number of search queries for the key word “Virtualization” in parallel to the increase in search queries for Cloud Computing.

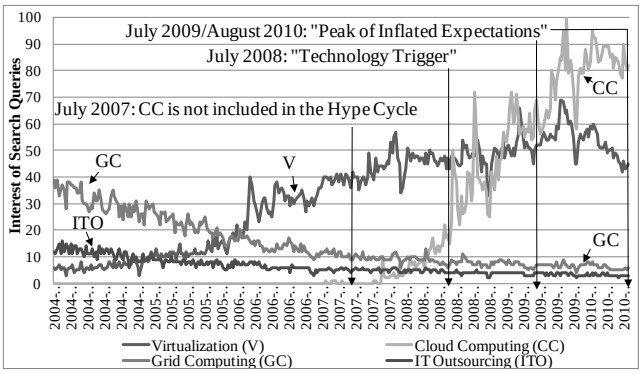

Figure 1: Search Queries for Cloud Computing and Related Concepts.

We compared our results with the technology hype cycles that are annually published by the Gartner Group [10] and integrated the

Submission to:
Wirtschaftsinformatik 2011, 16.-18. February, 2011, Zurich, Switzerland.

information from this source into Figure 1. The illustration encompasses two of the five phases “Technology Trigger”, “Peak of Inflated Expectations”, “Trough of Disillusionment”, “Slope of Enlightenment” and “Plateau of Productivity” [10]. Cloud Computing first appears at the phase of “Technology trigger” in the year 2008. In 2009 and 2010 it is assigned to the phase “Peak of Technology”, but with a superior maturity. The term “Private Cloud” was newly added in 2010 to the Gartner Hypecycle and was also assigned to the second phase, but close to its starting point. Gartner predicts a time span of two to five years until mainstream adoption [10].

It seems that the notion of “Cloud Computing” has been especially dominant in media aimed at readers with a practical background [23]. Mei et al. [25] regard academic discussions on research issues in Cloud Computing as being still inadequate. However, with the emergence of this new paradigm, research challenges come up that need to be adopted by the academic community [16]. New research opportunities emerge that may still be grounded in existing work on IT Outsourcing, IT Service Management as well as Risk and Compliance Management [8,18,23].

With this study we aim at gaining a better understanding of the growing and evolving Cloud Computing ecosystem, which encompasses a variety of business models, actors and market niches [26]. We analyze the ecosystem from both practical and academic perspectives and contrast these two different approaches. We attempt to identify the main concepts and actors that constitute the Cloud Computing ecosystem and also examine the obstacles and challenges associated with the adoption of this paradigm.

To achieve these research objectives, we adopt a quantitative content analysis approach [19,36]. We collected articles from practitioner-oriented outlets (magazine and internet articles) as well as scientific publications (articles published in scientific journals and conference proceedings) with a focus on Cloud Computing. Based on the literature, we identify major topics in Cloud Computing and evaluate them within the Cloud Computing ecosystem by means of positive and negative wordlists. In addition to the identified topics, we also analyze the significance of research challenges that are discussed in the literature. All these insights are finally brought together in a model of the Cloud Computing ecosystem that provides an overview of the main issues and main actors. The model is intended to further clarify the concepts, goals and motivations of Cloud Computing.

The article is structured as follows: Subsequently, related work is presented and discussed. In the third section we describe our chosen research method (quantitative content analysis) and provide details on the preprocessing phase, the process of analysis, and the used corpus. The results and main findings of our work are presented in the fourth section. Next, we discuss these findings in more detail and develop a model for the Cloud Computing ecosystem. Finally, we outline the limitations of our approach and give a brief summary.

2. RELATED WORK

Considering the general lack of a common definition of Cloud Computing [38], researchers have especially focused on gaining more insights into Cloud Computing and its multiple facets during the last few years. For instance, Youseff et al. [41] propose an ontology which illustrates the relevant components of Cloud Computing and their relationships. Researchers have also studied Cloud Computing with the aim of increasing the popularity of this research subject within the scientific community [4,18,41]. As yet, little research has been conducted on the drivers and actors of the Cloud Computing ecosystem, on the adoption of Cloud Computing services, or the success and risks associated with them [23]. Rather, existing studies on the emergence of new business models and the evolution of value chains were initiated because of new technological developments [18].

In view of the fact that Cloud Computing is mostly approached from a purely technical perspective, Leimeister et al. [18] extended the focus to include a broader understanding of business opportunities and business value. They describe the ongoing evolution from traditional IT Outsourcing towards Cloud Computing value networks.

Customers and providers are the main actors within these emerging Cloud Computing networks. Taking the customer perspective, Benlian [5] discussed the determinants for customer adoption of SaaS on the basis of transaction cost theory. He identified environmental uncertainty and application specificity as contributing factors for SaaS adoption. Koehler et al. [17] identified customer preferences for attributes of Cloud Computing services by means of choice-based conjoint analysis within an empirical study. They found that the average reputation of the Cloud Computing service provider and the use of standard data formats are more important than financial aspects such as cost reduction or pricing tariffs. Armbrust et al. [4] present a list of ten obstacles for Cloud Computing, of which the following three are considered as affecting adoption: availability/business continuity, data lock-in, and data confidentiality, and auditability. Although the forms of software delivery and pricing associated with Cloud Computing are assumed to replace some traditional software products in the long run, they are not expected to completely eliminate them in the near future [9].

From a vendor perspective, obstacles are identified that affect the growth of Cloud Computing as well as policy and business issues, e. g., data transfer bottlenecks [4]. Nevertheless, Cloud Computing facilitates the introduction of new products and services without large investments in IT infrastructure [31]. Pricing strategies and revenue models are suggested in order to exploit the economic opportunities of this emerging paradigm [3,31]. Huang and Wang [15] investigated the relationship between the SaaS software delivery model and the productivity of software vendors by examining 179 U.S. software companies. They identified demand-side diseconomies of scale for pure SaaS firms which make it difficult for them to compete with larger established software companies.

In view of the small number of studies that have dug deeper into the Cloud Computing ecosystem, there is a definite need for further research on this emerging research topic [16,25]. The goal of our study is to contrast the practical and the scientific view on Cloud Computing and to rigorously analyze the Cloud Computing ecosystem from both perspectives.

To the best of our knowledge, we are the first to apply quantitative content analysis to gain a holistic view on Cloud Computing that accounts for the arguments of both practice and academia. This approach allows us to draw a comprehensive picture of the issues that need to be tackled within this field as

well as of the opportunities it offers for research and practice alike.

3. QUANTITATIVE CONTENT ANALYSIS

Our approach constitutes a combination of term frequency and cluster analyses in the field of Cloud Computing. The general objective of a quantitative content analysis is to analyze, edit, and organize a corpus consisting of a set of documents to find hidden features and extract information for further processing [36]. Lijphart [19] stated that content analysis plays an important role for theory development in fields that still lack a theoretical background, as, for example, Cloud Computing.

Corpus: As sources of practice-related articles we chose the two IT magazines *CIO Magazine* and *MIT Technology Review*, as well as the two internet pages *Silicon.com* and *InformationWeek.com* which report regularly on the topic of Cloud Computing. Through the inclusion of both print and online publications we attempted to capture a wide range of topics. We excluded blog texts from our analysis due to the uncertain expertise of the authors and instead relied on the professional expert knowledge of the magazine and website editors. In view of the results of the Google search analysis we selected a time horizon from 2007 to August 2010 (cf. Figure 1). While these articles typically take a more subjective approach to their topics than peer-reviewed journal articles, they serve as a useful barometer of current practice and sentiment in the marketplace [24]. On the other hand, we conducted a systematic literature review of articles that appeared in scientific journals and the proceedings of information systems conferences. In our review, we applied keywords related to Cloud Computing (cloud, cloud computing, Software, Platform and Infrastructure as a Service, plus variants and abbreviations of these key words) and performed a forward and backward search in the indentified articles on Cloud Computing and related topics [44]. We searched the proceedings of the major international information systems conferences *ICIS*, *ECIS*, *AMCIS* and *HICSS a*s well information systems journals ranked by the Association for Information Systems (AIS) with ≤ 14.00 points [1] (cf. supplement: www.uwi.uos.de/supplementwi11.pdf). The identified Cloud Computing articles are categorized in Table 1.

Table 1: Description of the Corpus

Publication	Publication Type	# of Articles per Year				Overall # of Articles
		2007	2008	2009	2010	
CIO Magazine	Magazine	5	5	9	11	30
MIT Technology Review	Magazine	3	8	21	4	36
Silicon.com	Internet Articles	0	38	38	16	92
InformationWeek.com	Internet Articles	6	99	133	49	287
AIS Journal Ranking with ≤14	Scientific Journals	0	1	3	1	5
ICIS, ECIS, AMCIS, HICSS	Scientific Conferences	0	5	9	5	19
Other	Cited in Scientific Articles	1	6	3	6	16
Total		15	162	217	92	**485**

Software: The use of software for the quantitative content analysis is of particular importance because its capability of analyzing large volumes of data exceeds that of any human analyst. Another important benefit of using content analysis software tools is the consistency and reliability of the results [34]. We decided to apply the open source tool Rapidminer 5.0 and its text processing package. The advantage of this tool is its open source character which, in contrast to black-boxed systems, allows for customization [7].

Preprocessing: Before data processing could start, we copied the documents for analysis in text documents and deleted additional information like the reference list in scientific articles and text that came from online advertisement in practice-related articles. For the basic preprocessing of the documents, we followed a widely acknowledged information retrieval and text mining procedure applied by Sidorova et al. [35] and added an additional first stem operator that applies especially to Cloud Computing. One of the main problems of text analysis is the existence of search terms with different spellings. As Cloud Computing is an emerging, not highly matured topic [39], this problem is of particular significance. For example, during our analysis we found that the various existing spellings of the three "as a Service" types make it difficult to distinguish between them and impedes the process of analysis. We decided to summarize all "as a Service" spelling variants in the abbreviation "aas". This approach has the advantage that it captures all types of services in a bi-gram analysis. As the next main preprocessing steps, we transformed all words to lower case and tokenized the document into single terms. To clean this list, we deleted terms that have ≤ 2 tokens and applied a stopword list created by Loughran and McDonald [20]. This list contains currencies, dates, numbers, generic expressions like "and", "I" etc., names (first names and surnames), and places. Finally, we applied a stem list that we created for the top 50 words to consolidate words in singular and plural. The application of a stemmer like the Snowball or Porter stemmer stems words to close for our analyses [6]. For example, the word "cloudstack", which is the name of a Cloud Computing service, would be replaced by "cloud". Accordingly, such important differences between words are no longer visible. Finally, each word can be treated like a vector for further processing. The three types of analysis applied in this work are the counting of words, a document cluster analysis and an analysis of sentences that contain specific keywords. An overview of the analysis process is given in Figure 2, which depicts the analysis steps in chronological order.

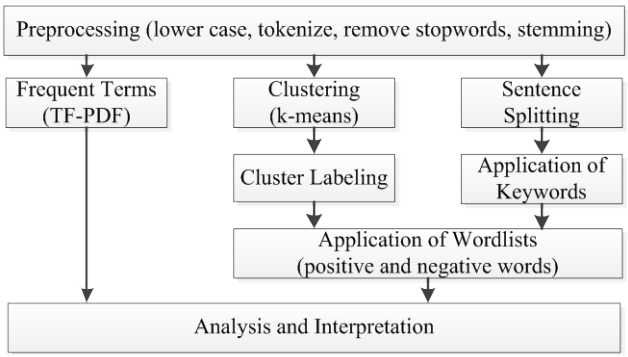

Figure 2: Process of Analysis

TF-PDF (weight of terms): To determine the significance of a term in a collection of documents, the term weighting scheme TF-IDF (term frequency – inverse document frequency) by Salton and

Buckley is often used in quantitative content analysis [34]. This algorithm assigns a large weight to terms that frequently appear in a single document, but rarely in a document collection. Thus, words that are usually assigned to a stopword list do not have high weights in this scheme. The aim of this weighting scheme is to retrieve documents that best match a search query. On the other hand, in our analysis we try to determine the so-called "hot topics" [7] in Cloud Computing. Hence, since the TF-IDF scheme is not adequate for our approach, we apply a modification, which is called TF-*PDF* (term frequency – proportional document frequency) [7]. In contrast to TF-IDF, the TF-PDF indicator applies an exponential instead of a logarithmic approach. Its calculation is shown in equation 1 with w_j as the weight of term j.

$$w_j = \frac{f_j}{F} \cdot \exp\left(\frac{n_j}{N}\right) \quad (1)$$

The first expression of the formula represents the term frequency, with f_j standing for the frequency of term j and F for the total number of terms in the entire corpus. In the second composition the exponential function is applied with n_j representing the number of documents that contain term j and N representing the total number of documents in the corpus. In our corpus, this method leads to an adjustment of the stopword list, because common words like "make" are listed in the results and need to be deleted. In summary, terms that occur in many documents are more helpful for the identification of main topics by means of TF-PDF. Furthermore, this algorithm has been validated in an experiment conducted by Bun and Ishizuka [7].

Clustering: For the identification of main topics in Cloud Computing we apply the clustering algorithm k-means by MacQueen [21]. This non-hierarchical cluster analysis with square Euclidean distances assigns every document to one particular cluster. It needs to be mentioned that this algorithm uses a heuristic approach, which means that the global optimum will not be reached in every process. We decided for this algorithm, since it is commonly known, works very efficient (it needs little computing power) and works with several types of data [14]. The number of clusters needs to be determined by the user. We use an approximation approach [22] which is based on the number of documents (n) as shown in equation 2.

$$k \approx \left(\frac{n}{2}\right)^{0,5} \quad (2)$$

The four main steps of the algorithm are as follows [14]: Firstly, k arithmetic means are randomly selected. Secondly, k clusters are created by assigning the documents to the nearest neighbor of the k centroids (cluster prototype). Thirdly, new centroids are calculated on the basis of the new allocation of documents. This step is repeated until the centroids stop changing. The cluster labels are developed inductively by logically reviewing main keywords which are here called centroids [6,11].

Sentiment Analysis: Finally, we applied word lists containing terms with either positive (e. g. "benefit", "desired") or negative (e. g., "interrupt", "mistake") connotations [28]. These word lists were developed by Loughran and McDonald [20] who applied terms from the Harvard Psychosociological Dictionary (Harvard-IV-4) to the field of business and economics. The main difference between the Harvard list and the list by Loughran and McDonald lies in the connotations assigned to certain terms. For example, "cost" and "capital" are categorized as negatively associated words on the Harvard list, but are discussed in business and economics on a neutral basis. In order to apply the Loughran-McDonald list to the field of information systems, some minor adjustments were necessary.

4. ANALYSIS

4.1 N-Gram Analyses

We analyzed the data from our two corpuses separately. They were transferred into numerical vectors of word frequencies. Each position in the vector corresponds to a single word (uni-gram) in the corpus [42]. For each corpus, we determined the 25 most influential terms. We considered this number of terms to provide a representative depiction of the current discussions on Cloud Computing. The results of the uni-gram analyses are shown in Table 2. The top 10 to 15 terms are almost similar in both lists. However, taking a closer look, there are also recognizable differences. As regards the practice publications, technical issues and market actors seem to be the most dominant themes. Terms like "technology", "storage", "server", "software" and "platform" point at the frequent discussions centered on the technical implementation of Cloud Computing. A lot of discussions also focus on large vendors in the Cloud Computing market, as e. g. Microsoft, Google and Amazon. Security is another key term that was identified in the analysis of practitioner-oriented publications

Table 2: Top 25 Uni-Gram Ranked by TF-PDF

Practice		Science	
Term	**TF-PDF**	**Term**	**TF-PDF**
cloud	0.09990	service	0.06699
computing	0.04328	cloud	0.05800
company	0.04182	computing	0.04128
service	0.03938	customer	0.03478
application	0.03727	application	0.02930
customer	0.02521	resource	0.02587
data	0.02349	vendor	0.02354
business	0.01669	data	0.02294
software	0.01618	company	0.02017
vendor	0.01416	model	0.01992
server	0.01404	business	0.01635
system	0.01124	system	0.01506
technology	0.00858	software	0.01230
web	0.00857	management	0.01215
microsoft	0.00842	grid	0.01176
security	0.00797	server	0.01143
amazon	0.00774	cost	0.01134
google	0.00759	infrastructure	0.01019
center	0.00726	time	0.00873
infrastructure	0.00639	technology	0.00808
cost	0.00623	web	0.00800
management	0.00621	process	0.00764
platform	0.00512	information	0.00762
storage	0.00504	storage	0.00698
time	0.00496	saas	0.00594

In general, researchers tend to use a similar vocabulary when discussing Cloud Computing. However, instead of using concrete terms like "server" and "storage" they prefer abstractions like "resource" and "system". The term "grid" is frequently mentioned, for Grid Computing is regarded by many as one of the predecessors of Cloud Computing, and both concepts are often directly compared to each other [39]. Also, service-related issues seem to be more prevalent in academic publications on Cloud

Computing, as apparent in the frequent use of the terms “service” and “saas”. Moreover, the occurrence frequency of the terms “business” and “cost” suggests that scientific articles often discuss the effects of Cloud Computing on companies.

The initial search focused only on single words. In a second step, we extended our search to bi-gram analyses, again for each corpus separately. The objective is to gain a deeper understanding of compounded words. Bi-grams consist of exactly two consecutive words [42]. The following results show considerably lower TF-PDF values than those of the uni-gram analyses (cf. Tables 2 and 3). This is the case because recurrences of the same two-word sequence (e. g., “cloud_computing” and “cloud_service”) are less frequent compared to a single word (e. g., “cloud”).

Again, there are striking analogies between practice-oriented and scientific publications. In both lists, the bi-grams „cloud_computing“, „data_center“, and „cloud_service“ belong to the top three combinations. In the practice corpus, the term “cloud” is more often part of word combinations than in the scientific corpus. Moreover, Amazon’s service “Elastic Compute Cloud” (also called „EC2“) is mainly discussed among practitioners, as can be derived from the frequent occurrence of the bi-grams “[elastic] compute_cloud” and “amazon_ec” [2].

Table 3: Top 25 Bi-Grams Ranked by TF-PDF

Practice		Science	
Term	**TF-PDF**	**Term**	**TF-PDF**
cloud_computing	0.03286	cloud_computing	0.01991
data_center	0.00550	data_center	0.00445
cloud_service	0.00343	cloud_service	0.00430
private_cloud	0.00205	service_vendor	0.00360
virtual_server	0.00168	cloud_vendor	0.00243
cloud_vendor	0.00166	web_service	0.00240
open_source	0.00162	virtual_server	0.00205
web_service	0.00141	grid_computing	0.00195
software_aas	0.00140	business_model	0.00160
google_application	0.00116	business_process	0.00131
service_vendor	0.00110	computing_cloud	0.00130
public_cloud	0.00109	computing_resource	0.00123
operating_system	0.00109	cloud_application	0.00108
web_application	0.00100	computing_service	0.00107
computing_service	0.00080	application_service	0.00091
amazon_web	0.00079	service_level	0.00088
cloud_application	0.00072	utility_computing	0.00083
amazon_ec	0.00065	service_delivery	0.00080
company_cloud	0.00065	service_computing	0.00076
application_cloud	0.00060	operating_system	0.00074
end_customer	0.00053	economies_scale	0.00071
public_sector	0.00052	knowledge_area	0.00070
application_service	0.00051	resource_management	0.00069
compute_cloud	0.00051	pricing_model	0.00067
saas_application	0.00048	software_aas	0.00067

Aspects of service provision are again more prevalent in the scientific corpus. In contrast to the practitioner outlets, scientific publications often deal with the management and adoption of Cloud Computing services within companies, as exemplified by the frequent use of terms like “service_level”, „business_model“, “service_delivery” and „business_process”. Surprisingly, the term “economies_[of]_scale” is one of the top 25 terms already. Thus, there might be a first tendency towards the study of theories related to the Cloud Computing phenomenon.

4.2 Cluster Analyses

The main objective of the cluster analysis is to assign the documents of each corpus to the most frequently discussed themes. The three obligatory parameters for this algorithm are the maximal numbers of runs and the maximal optimization steps. The first parameter defines the number of runs with a random initialization for the first centroid, which we set 10. The maximal optimization steps define the number of iterations performed for one run of the algorithm, which we set 100. We determined the number of clusters for each corpus with the presented approximation approach (cf. equation 2) [22], resulting in 15 clusters for the practitioner corpus (445 documents) and 5 clusters for the academic corpus (40 documents). Due to the heuristic nature of the k-means algorithms, minor deviations occurred with regard to documents brought together in clusters. As a consequence, we conducted the analyses several times. Two authors of this paper subjectively decided on the most adequate result to serve as the basis for these analyses.

For presentation and discussion of the results, we sorted the topic clusters descending by the number of documents they include (cf. Table 4 and Table 5). The clusters were labeled by means of logical reviewing [6]. To improve the quality of labels, again, two authors of this paper were involved in independently reviewing and coding the results of the cluster analysis.

Additionally, we conducted an analysis of positive and negative words which were used in the individual cluster documents (in the supplement, we present a list of centroids with a factor loading ≥ 0.05 of each cluster: www.uwi.uos.de/supplementwi11.pdf). Here, we would like to mention that the outcomes need to be interpreted with care. They represent the sentiment of the entire cluster and cover every sentence of each cluster document. Thus, there could be a bias, which we address in the following sentiment analyses (cf. section 4.3). Nevertheless, the results indicate first sentiment tendencies.

As for practitioner articles (cf. Table 4), it becomes obvious that “General Topics”, “Technical Topics” and “Company Perspective (Cloud Computing)” are the three most prevalent clusters. The sentiment analysis revealed that in the discussion of general topics more positive than negative words are used. A more pessimistic view is taken on technical issues, which is partly due to the still maturing interface and architecture concepts. Another interesting aspect is that similar topics are covered by different clusters; this is true, for example, for clusters 3 and 4. The articles that belong to these clusters use different vocabularies and therefore express different sentiments. The articles of cluster 3 embrace vocabularies that are used within the context of Cloud Computing with a balanced sentiment. In cluster 4, most of the terms are closely related to the topic of IT Outsourcing, with which a wider range of practitioners is already familiar [18], and more positive than negative words are used. Several clusters refer to the main actors in the current market, as e. g., Microsoft, Amazon Web Services (AWS), Nasa (Nebula), Oracle, Salesforce (covered in cluster 5), as well as open source products and services. Different vendors have different reputations on the market, whereas in this respect, mature services are usually in a better position (as, for example, AWS). Also, risk and security issues are obviously much debated. This becomes evident when looking at cluster 12, which contains only documents that exclusively deal with this field. The sentiment in this field is slightly positive, since all words in the

cluster are considered (The discussion of this result is presented in section 4 and 5). Finally, there is a small cluster comprising three documents about IT Outsourcing and the Cloud Computing market.

Table 4: Results of the Cluster Analysis (Practice)

#	Cluster	# of Documents (Percentage)	Positive Words	Negative Words
1	General Topics	92 (20.7%)	**58.9%**	41.1%
2	Technical Topics	64 (14.4%)	41.1%	**58.9%**
3	Company Perspective (Cloud Computing)	54 (12.1%)	48.9%	**51.1%**
4	Company Perspective (IT Outsourcing)	37 (8.3%)	**56.4%**	43.6%
5	SaaS (Provider)	31 (7.0%)	**65.6%**	34.4%
6	Microsoft Azure	31 (7.0%)	38.1%	**61.9%**
7	Vendors	27 (6.1%)	**56.3%**	43.8%
8	SaaS (Business/ Management)	23 (5.2%)	41.1%	**58.9%**
9	Government	21 (4.7%)	**72.9%**	27.1%
10	Open Source/ Standards	20 (4.5%)	**57.3%**	42.7%
11	Amazon Web Services	20 (4.5%)	**70.5%**	29.5%
12	Security/ Risk	10 (2.2%)	**54.3%**	45.7%
13	Nasa Nebula	7 (1.6%)	31.0%	**69.0%**
14	Oracle Fusion	5 (1.1%)	**56.0%**	44.0%
15	IT Outsourcing/ Cloud Computing Market	3 (0.7%)	46.4%	**53.6%**
	Overall	445 (100.0%)	**48.7%**	51.3%

The analysis of scientific articles resulted in a categorization into five clusters (cf. Table 5). Again, the major cluster comprises articles on general topics from the field of Cloud Computing, showing positive attitudes. The second cluster consists of literature on resource management of Cloud Computing services in which slightly more negative than positive words are used. This cluster is strongly dominated by researchers like Püschel et al. (for example [30]). Topics regarding Grid vs. Cloud Computing are addressed in the articles of cluster 3, which shows a strong positive sentiment. The fourth cluster is dedicated to issues concerning sourcing models like SaaS and classic IT Outsourcing. Here, the basic sentiment of the articles is positive. Finally, there is the fifth cluster that consists of articles with topics on implications for business and management with a strong positive sentiment. This might be due to researchers that discuss and develop concepts and methods for simplifying business processes and reducing costs by means of Cloud Computing services.

Table 5: Results of the Cluster Analysis (Science)

#	Cluster	# of Documents (Percentage)	Positive Words	Negative Words
1	General Topics	16 (40.0%)	42.5%	**57.2%**
2	Resource Management	8 (20.0%)	46.5%	**53.5%**
3	Grid vs. Cloud Computing	8 (20.0%)	**67.2%**	32.8%
4	SaaS/ IT Outsourcing	4 (10.0%)	**55.1%**	44.9%
5	Business/ Management	4 (10.0%)	**68.8%**	31.2%
	Overall	40 (100.00%)	**52.4%**	47.6%

The analysis of scientific articles proved to be a lot more challenging than the review of practitioner-oriented publications. In comparison, after preprocessing, the 40 analyzed scientific articles contained 104,222 single terms whereas the 445 practice-related articles contained 158,121 single terms. Thus, assigning a scientific article to one particular cluster caused difficulties. The results presented in Table 5 show that by and large, only a handful of major research topics can be currently distinguished in the field of Cloud Computing. All topics outside these main categories are usually discussed in the context of overview articles. The results call for further in-depth analyses of these articles.

4.3 Sentiment Analysis

The cluster analysis helped to identify major topics in Cloud Computing, while the sentiment analysis revealed a first trend of opinions in the field. However, a deeper understanding of positive and negative sentiments was still lacking. Therefore, a further sentence analysis was conducted which consisted of several processing steps. Firstly, sentences were split up by identifying punctuation marks. Within these sentences we searched for keywords covering particular topics and drivers of Cloud Computing. Finally, we marked positive and negative terms to make them countable.

Table 6: Major Topics in Cloud Computing

Topic/Description	Concepts (Synonyms)
Technology - Changing requirements for IT infrastructures and architectures [16] - Resource management (virtualization and the absorption of demand peaks) [4,16,37] - Standardization of interfaces [16]	hardware, server, virtual, resource, infrastructure, network, middleware, rout, center, interface, storage
Costs - Cost management (cost for migration, allocation of costs, cost savings) [16,18] - Pricing models for Cloud Computing Services [16] - Implementation and consulting costs [18]	budget, pric, money, cost, accounting, acountanc, finance, saving, save, pay, tco
Personnel - Changing role of IT department and political implications on (IT) personnel [16] - Effects on end users [16,37]	skill, personnel, fluctuation, manpower, workforce, labor, employee, user, department, staff
Security - Security issues: denial of service attacks, threats, malware [32,37,39] - privacy issues: data protection and treatment [4,16]	protection, hacker, secur, recover, confidential, property, privacy, vulnerabilit, delet, threat, trust, privacy, denial, Malware, unauthoriz, risk
Quality - Service availability and business continuity [37,39] - Elasticity (Resilience) and performance [4,16]	performance, availab, quality, assurance, iso 9000, six sigma, dependability, resilience, requirement, stability, stable, continu, elastici, flexib
Compliance - Regulatory requirements that restrict data movement and processing [29,39] - Ability to audit Cloud Computing services [16,39]	regulat, law, government, liability, penalt, legislation, rule, legal, compliance, jurisdiction, licens, audit

Roberts [33] points out that the results of a content analysis always need to be interpreted within the general context of the research field to determine the full meaning of a particular term. Even the selection of cluster labels needs to be theoretically underpinned. Thus, to explore the Cloud Computing ecosystem systematically, we developed a list of drivers and factors on the basis of scientific literature which was assigned to the first cluster (General Topics) of the scientific corpus (cf. Table 5). Some of these articles contain discussions about open issues in Cloud Computing and suggest research agendas, which were merged into 6 key topics as described in Table 6. Also, we added concepts that could be used as synonyms for the analysis. These concepts were

derived from the results of the n-gram analyses. In order to be able to detect different word forms of the same word stem (e. g. plural and singular terms; nouns and adjectives) we shortened the words to their stem where needed (e. g., "secur" instead of "secure" and "security").

The results of the sentiment analysis on the basis of particular sentences are presented in Table 7. We ranked the topics by the TF-PDF factors of the practitioner corpus, which are quite similar to those of the scientific one.

Table 7: Results of the Sentiment Analyses

Topic	Practice			Science		
	TF-PDF	Positive Words	Negative Words	TF-PDF	Positive Words	Negative Words
Technology	0.0659	**54.4%**	45.6%	0.0822	**58.8%**	41.2%
Costs	0.0186	**51.7%**	48.3%	0.0262	**57.1%**	42.9%
Personnel	0.0177	46.7%	**53.3%**	0.0225	49.9%	**50.1%**
Security	0.0143	30.2%	**69.8%**	0.0095	29.9%	**70.1%**
Quality	0.0087	**53.7%**	46.3%	0.0198	48.7%	**51.3%**
Compliance	0.0056	44.3%	**55.7%**	0.0049	41.6%	**58.4%**

In contrast to the results of Table 4, the outcomes presented here show a different picture of particular topics. Main causes are discussed in the subsequent section (cf. section 5). However, technological issues are seen positive by both practitioners and researchers. Interesting is the difference for cost issues. Researchers discuss cost issues in Cloud Computing more positively than practitioners. The most significant outcome is the strong negative sentiment in sentences that comprise expressions of security issues. The opinion on quality varies slightly different between both groups. Finally, compliance topics reveal as well as security topics a rather negative connotation.

5. DISCUSSION OF RESULTS

Exploring the Cloud Computing ecosystem from different perspectives offers interesting insights into the discrepancy between science and practice. For instance, the n-gram and cluster analyses revealed a strong focus on Cloud Computing providers in practice (cf. Table 3). Obviously, user companies are interested in new Cloud Computing services and products. Especially popular and long-established providers (like AWS and Salesforce) have a positive reputation (cf. Table 4), as they were first movers in Cloud Computing. In contrast, Microsoft's development platform Azure is discussed less benevolently (61.9 % negative words).

The topic "technology" receives quite a positive interpretation in both practice and science (cf. Table 7). In comparison to Table 4 in which technical issues are evaluated rather negatively, a more detailed analysis is necessary. For instance, researchers [12] wrote:"A key concept in cloud computing is that cloud providers can use **resources* more **pos*efficiently* through statistical multiplexing, and may operate at lower cost than medium-sized data centers" (words that match the topic are highlighted with a "*"; positive/negative words by "*pos*" or "*neg*"). In practitioner-oriented articles, sentences can be found like:"Scaling a web application – adjusting **resources *pos*smoothly* in response to growing traffic – is a do-or-die proposition for most web startups."[27] However, the analysis of cluster 2 "Technical Topics" (cf. Table 4) reveals that in the respective articles expressions like "problem", "costly" and "difficult" are used frequently, leading to a slightly negative sentiment (58.9% negative words). Nevertheless, we assume that the sentence-based sentiment analysis (Table 7) provides a more reliable picture on technical topics.

Security issues in Cloud Computing offer interesting results as well. Table 4 and 5 suggest that security is positively discussed in practice. The outcomes presented in Table 7 provide a contradicting impression. In both practice and science, security issues are discussed fairly negatively. Here, the question arises, why there is no cluster which deals with security topics in science. Of course, several authors touch security issues, but their works on this topic are by far not as comprehensive so that the cluster algorithm could shape an additional cluster. For instance, some articles represent research in progress [32] and others are largely restricted to mere descriptions of the Cloud Computing paradigm. Moreover, an analysis of the term "security" shows that the strongest influence in science is shown in the general topic cluster (centroid: 0.046). Summarized, security issues are recognized as a success factor for Cloud Computing in both science and practice, but a strong research field is not built yet.

Another negatively associated topic is compliance, which is exemplified by the following sentences from the scientific corpus:"From an individual's perspective, cloud computing presents **neg*risks* of personal data exposure, and **neg*lack* of awareness regarding the location and **jurisdiction* of their data."[16] On the other hand, the following sentence is typical for a practitioner-based article:"For example, if there's a security **failure* in a service that comprises financial data, a company might be required to notify customers under state or federal **law*, and potentially face legal action."[2]

The discussion on data centers (which are occasionally called clouds [40]) points at another difference and is worth discussing. It becomes evident from Table 3 that practitioners frequently discuss the topic of "clouds". In general, cloud concepts are differentiated between private (internal), public (external) and hybrid (hybrid types of the aforementioned) clouds [4]. In science, this topic is not extensively discussed (cf. Table 2 and 3). For example, Wlodarczyk et al. [40] support this finding as well and provide a first insight by developing an inter-company solution to deal with security issues.

Summing up, in both practice and science there seems to be a detailed discussion what Cloud Computing actually is and is not [40]. In science the tone is slightly more negative on general topics, but in the end Cloud Computing has a quite positive sentiment. The three negative associated topics security, compliance and personnel indicate open issues. Apparently, companies have problems in adopting Cloud Computing services and integrate them into their IT architecture. Researchers try to uncover the core of Cloud Computing by analyzing business models and business processes (cf. Table 3), while practitioners are more interested in revealing information about market actors and new Cloud Computing services.

Figure 3 gives an overview about the Cloud Computing ecosystem as resulting from our qualitative content analyses. It synthesizes the major topics and most relevant key words related to the still evolving Cloud Computing paradigm. Words discussed only in practice are highlighted with a "*". Purely scientific notions are marked by a "†". All other words are relevant to both practice and academia.

We structured the topics and key words along the Cloud Computing service process from provider to customer. The stakeholders (e. g., provider and customer) act on the basis of legal and compliance requirements as depicted by the Government/ Compliance box. The provision of Cloud Computing services is related to technical issues. Security issues and risks affect stakeholders and the provision of services. They are also linked to the technical issues.

6. LIMITATIONS

The applied research method (quantitative content analysis) and the design of this study imply some unavoidable limitations. One major problem lies in the interpretation of word lists. Software tools are unable to differentiate between different meanings of the same word [43]. Therefore, in some cases, false negatives or positives might have been included into the analysis. In response to this problem, we tried to follow the recommendations of Roberts [33] by providing a theoretical basis for our cluster analysis and by putting it in the context of the overall debate on Cloud Computing.

It is also important to note that, in addition to practitioner-based publications, our corpus predominantly comprises North American high quality scientific journals which are included in the AIS ranking. One may argue that the scope of our analysis was critically limited by this approach. However, with our study we intended to identify the main differences between current scientific and practical understandings of Cloud Computing. The inclusion of additional sources which are closer to one of the corpuses in terms of domain affiliation and word usage could have led to fuzzy results. Also, we focused on North American sources because from our point of view, the main driving forces behind Cloud Computing are still to be found in North America. Differences between North American and European research that are commonly acknowledged need to be considered [35]. These limitations must be kept in mind when interpreting the results of our analyses.

In addition, our way of labeling the clusters may have been subject to biases. However, we did our best to minimize this risk by carefully examining term loadings and by having the clusters labeled by two researchers independently [35].

Finally, the choice of the k-means cluster algorithm entails some limitations, too. We could have applied several other algorithms or improvements of k-means [14], but decided against it because of the efficiency and widespread familiarity of the k-means cluster algorithm.

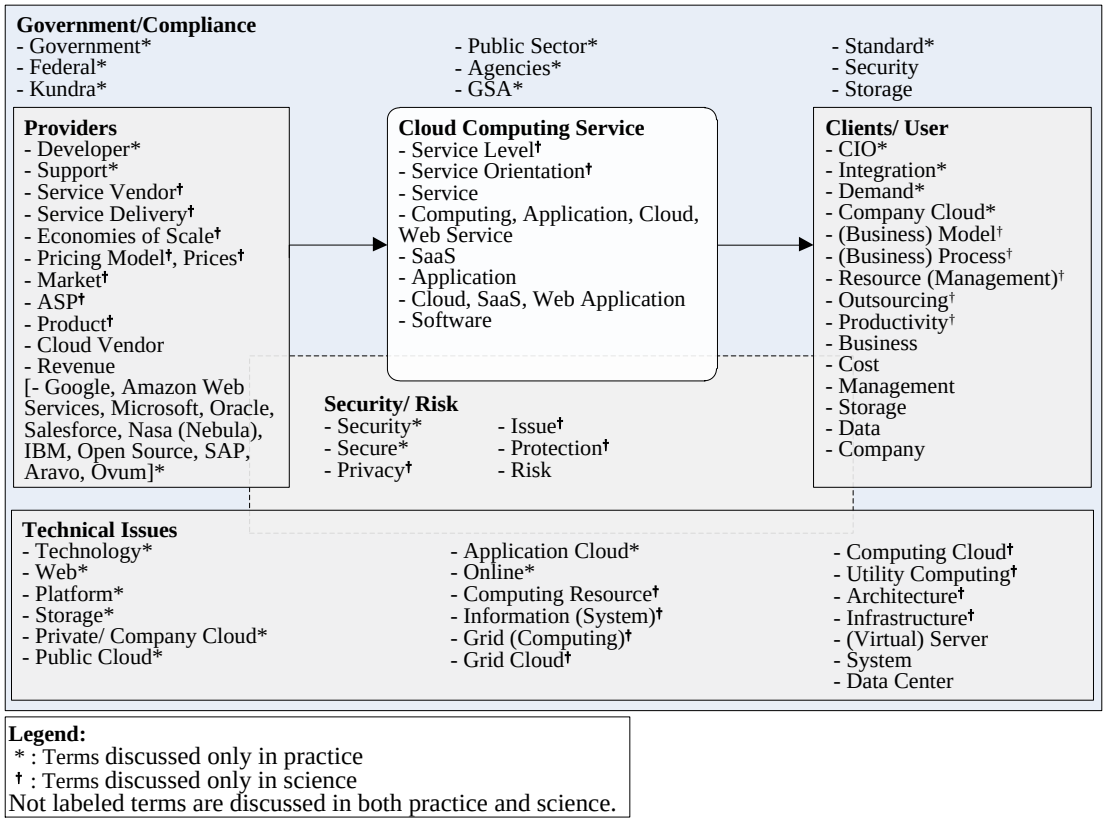

Figure 3: Cloud Computing Ecosystem

7. CONCLUSIONS

In this paper, we explored the Cloud Computing paradigm from both a scientific and practitioner-based perspective by applying quantitative content analysis. The contribution of this paper is twofold: First, it identifies the key terms and topics that are part of the current Cloud Computing discussion in practice and academia. We aggregated the key terms and topics into a model of the Cloud Computing ecosystem. This model reflects the overall results in the form of a simple Cloud Computing service process (see Figure 3). Second, this paper discloses the sentiments of key topics as reflected in articles from both corpuses. Here, major findings are that Cloud Computing is seen positively in general. There are only few topics that practitioner-oriented outlets and academics evaluate rather negative. Results of the sentiment analyses vary between practice and science.

It is important to keep in mind that this research approach has its limitations. However, we tried to minimize biases by following a well established research approach. We are confident that our corpuses provide a high level of quality and are suited for the distinction between practice and science.

Due to the fast moving Cloud Computing market we are aware of our results being transient. Nevertheless, we hope that the outcomes of our study can be practically used to help researchers align their research topics to business needs and position their research topics within the Cloud Computing ecosystem. For future research we imagine that a bilingual study (German and English) of similar design could reveal deeper insights in geographical and cultural differences within the global discussion on Cloud Computing.

8. REFERENCES

[1] AIS. MIS Journal Rankings. Retrieved on 2010-08-23 from http://ais.affiniscape.com/displaycommon.cfm?an=1&subarticlenbr=432.

[2] Adam, E., Berlind, D., Hoover, J.N., and Foley, J. 2008. A How-To Guide To Cloud Computing. *InformationWeek.* Retrieved on 2010-08-23 from http://www.informationweek.com/news/services/storage/showArticle.jhtml?articleID=212201920.

[3] Anandasivam, A. and Premm, M. 2009. Bid price control and dynamic pricing in clouds. *European Conference on Information Systems (ECIS)*, (Verona, Italy, 2009).

[4] Armbrust, M., Fox, A., Griffith, R., et al. 2010. A view of cloud computing. *Communications of the ACM 53*, 4 (2010), 50-58.

[5] Benlian, A. 2009. A transaction cost theoretical analysis of Software-as-a-Service (SaaS)-based sourcing in SMBs and enterprises. *European conference on Information Systems (ECIS)*, (Verona, Italy, 2009).

[6] Blake, R. 2010. Identifying the core topics and themes of data and information quality research. *Americas Conference on Information Systems (AMCIS)*, (Lima, Peru, 2010).

[7] Bun, K. and Ichizuka, M. 2006. Emerging topic tracking system in WWW. *Knowledge-Based Systems 19*, 3 (2006), 164-171.

[8] Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., and Brandic, I. 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems 25*, 6 (2009), 599-616.

[9] Cusumano, M. Cloud computing and SaaS as new computing platforms. *Communications of the ACM 53*, 4 (2010), 27.

[10] Feen, J. 2009. Emerging Technology Hype Cycle 2010: What's Hot and What's Not. *Gartner*. Retrieved on 2010-08-23 from http://www.gartner.com/it/content/1395600/1395613/august_4_whats_hot_hype_2010_jfenn.pdf.

[11] Griffiths, T. and Steyvers, M. Finding scientific topics. *Colloquium of the National Academy of Sciences*, (2004).

[12] Günther, O., Müller, C., and Ziekow, H. 2010. RFID in the Cloud: A Service for High-Speed Data Access in Distributed Value Chains. *Americas Conference on Information Systems (AMCIS)*, (2010).

[13] Hof, R.D. 2006. Jeff Bezos' Risky Bet. *BusinessWeek.* Retrieved on 2010-08-23 from http://www.businessweek.com/magazine/content/06_46/b4009001.htm.

[14] Hotho, A., Nürnberger, A., and Paaß., G. 2005. A Brief Survey of Text Mining. *Journal for Computational Linguistics and Language Technology 20*, 1 (2005), 19-62.

[15] Huang, K. and Wang, M. 2009. Firm-Level Productivity Analysis for Software as a Service Companies. *International Conference on Information Systems (ICIS)*, (Phoenix, AZ, 2009).

[16] Khajeh-Hosseini, A., Sommerville, I., and Sriram, I. 2010. Research Challenges for Enterprise Cloud Computing. 2010. Retrieved on 2010-08-23 from http://arxiv.org/pdf/1001.3257.

[17] Koehler, P., Anandasivam, A., and Dan, M.A. 2010. Cloud Services from a Consumer Perspective Cloud Services from a Consumer Perspective. *Americas Conference on Information Systems (AMCIS)*, (Lima, Peru, 2010).

[18] Leimeister, S., Riedl, C., Böhm, M., and Krcmar, H. 2010. The Business Perspective of Cloud Computing: Actors, Roles, and Value Networks. *European Conference on Information Systems (ECIS)*, (Pretoria, South Africa, 2010).

[19] Lijphart, A. 1971. Comparative Politics and the Comparative Method. *American Political Science Review 65*, September 1971 (1971), 682-693.

[20] Loughran, T. and McDonald, B. 2010. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, Forthcoming.

[21] MacQueen, J.1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, 281-297.

[22] Mardia, K.V., Kent, J.T., and Bibby, J.M. 1997. *Multivariate analysis*. Acad. Press.

[23] Martens, B. and Teuteberg, F. 2009. Why Risk Management Matters in IT Outsourcing - A Systematic Literature Review and Elements of a Research Agenda. *European Conference on Information Systems (ECIS)*, (Verona, Italy 2009).

[24] McLaughlin, D. and Peppard, J. 2006. IT backsourcing: from 'make or buy' to 'bringing IT back in-house'. *European Conference on Information Systems (ECIS)*, (Göteborg, Sweden, 2006).

[25] Mei, L., Chan, W., and Tse, T. 2008. A Tale of Clouds: Paradigm Comparisons and Some Thoughts on Research Issues. *IEEE Asia-Pacific Services Computing Conference*, (2008), 464-469.

[26] Mell, P. and Grance, T. 2009. NIST Definition of Cloud Computing. *National Institute of Standards and Technology, Information Technology Laboratory*. Retrieved on 2010-08-23 from http://csrc.nist.gov/groups/SNS/cloud-computing/.

[27] Naone, E. 2008. Reaching for the Clouds. *MIT Technology Review*. Retrieved on 2010-08-23 from http://www.technologyreview.com/business/21127/.

[28] Pang, B., Lee, L., Rd, H., and Jose, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *ACL Conference on Empirical Methods in Natural Language Processing*, (2002), 79-86.

[29] Pearson, S. 2009. Taking account of privacy when designing cloud computing services. *ICSE Workshop on Software Engineering Challenges of Cloud Computing*, (2009), 44-52.

[30] Pueschel, T. and Neumann, D. 2009. Management of cloud infrastructures: Policybased revenue optimization. *International Conference on Information Systems (ICIS)*, (Phoenix, AZ, 2009).

[31] Pueschel, T., Anandasivam, A., Buschek, S., and Neumann, D. 2009. Making money with clouds – Revenue optimization through automated policy decisions. *European Conference on Information Systems (ECIS)*, (Verona, Italy, 2009), 1-13.

[32] Ramireddy, S., Chakraborthy, R., and Raghu, T. 2010. Privacy and Security Practices in the Arena of Cloud Computing-A Research in Progress. *Americas Conference on Information Systems (AMCIS)*, (Lima, Peru, 2010).

[33] Roberts, C.W. 2000. A Conceptual Framework for Quantitative Text Analysis. *Quality & Quantity 34*, 3 (2000), 259-274.

[34] Salton, G. and Buckley, C. 1988. Term-weighting approached in automatic text retrieval. *Information Processing and Management 14*, 5 (1988), 513-523.

[35] Sidorova, A., Evangelopoulos, N., Valacich, J.S., and Ramakrishnan, T. 2008. Uncovering the intellectual core of the information systems discipline. *MIS Quarterly 32*, 3 (2008), 467-482.

[36] Sullivan, D. 2001. *Document Warehousing and Text Mining*. Wiley Computer Publishing.

[37] Vaquero, L.M., Rodero-Merino, L., Caceres, J., and Lindner, M. 2009. A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review 39*, 1 (2009), 50-55.

[38] Wang, L. and von Laszewski, G. 2008. Scientific cloud computing: Early definition and experience. *IEEE International Conference on High Performance Computing and Communications*, (2008), 825-830.

[39] Weinhardt, C., Anandasivam, A., Blau, B., et al. 2009. Cloud Computing – A Classification, Business Models, and Research Directions. *Business & Information Systems Engineering 1*, 5 (2009), 391-399.

[40] Wlodarczyk, T.W., Rong, C., and Thorsen, K.A. 2009. Industrial Cloud: Toward Inter-enterprise Integration. *Lecture Notes in Computer Science 5931*, (2009), 460–471.

[41] Youseff, L., Butrico, M., and Da Silva, D. 2008. Toward a Unified Ontology of Cloud Computing. *2008 Grid Computing Environments Workshop*, (2008), 1-10.

[42] Zhang, T. and Oles, F. 2001. Text Categorization Based on Regularized Linear Classification Methods. *Information Retrieval 4*, 1 (2001), 5-31.

[43] Zhou, Y., Fleischmann, K.R., and Wallace, W.A. 2010. Automatic Text Analysis of Values in the Enron Email Dataset: Clustering a Social Network Using the Value Patterns of Actors. *Hawaii International Conference on System Sciences*, (2010).

[44] vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., and Cleven, A. 2009. Reconstructing the giant: on the importance of rigour in documenting the literature search process. *European Conference on Information Systems (ECIS)*, (Verona, Italy, 2009), 2206-2217.