IMPACT OF DICTIONARIES ON AUTOMATED CONTENT ANALYSIS – THE USE OF COMPOUND CONCEPTS IN ANALYSING KNOWLEDGE MANAGEMENT RESEARCH

Complete Research

Nora Fteimi, University of Passau, Passau, Germany, nora.fteimi@uni-passau.de

Dirk Basten, University of Cologne, Cologne, Germany, basten@wiso.uni-koeln.de

Abstract

Within the knowledge management (KM) domain, we perceive an increasing number of publications. Considering this increase, content analysis (CA) is a popular and empirically established method to handle context-sensitive information and to achieve an improved understanding. While CA can be applied in an automated way by using software solutions, a problem concerns the analysis of compound concepts (e.g. "intellectual capital"). Whereas dictionaries (i.e. lists of compound concepts) have been suggested to solve this problem, lack of research exists concerning the impact of using such a dictionary. By focusing on the KM domain and using the example of 614 publications within the Journal of Knowledge Management (JoKM), this paper aims to evaluate the impact of dictionaries for automated CA. We perform CA applying the automated approach with and without using a self-developed KM dictionary. The results were compared in terms of result similarity as well as result relevance to the KM discipline. Our findings reveal that using a dictionary for automated CA can lead to an improved context understanding and time savings. However, these benefits are opposed by subjectivity that results from the manual extraction of compound concepts to be used in the dictionary.

Keywords: Automated content analysis, compound concepts, knowledge management.

1 Introduction

Numerous social and technological developments had and have significant impact on knowledge creation. In recent years, there is an enormous increase of information and knowledge in form of textual information (Dwivedi and Venkitachalam 2009). To name a prominent example, "Wal-Mart handles more than 1 m[illion] customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America's Library of Congress" (The Economist 2010, p. 4). In view of publications within the information systems (IS) domain, there is, for instance, a major increase in publications within the journals of the Senior Scholars' Basket of Eight over the past decades (e.g. 1,129 research articles between 2000 and 2006 vs. 1,620 research articles between 2007 and 2013).

A challenge arises concerning a common understanding, which can be exemplified in the knowledge management (KM) domain (Serenko et al. 2010; Heisig 2009; Vorakulpipat and Rezgui 2008). As a young and interdisciplinary subdomain within IS research (Dalkir 2013), KM holds a crucial and established role in scientific research (Holsapple and Wu 2008). Related publications comprise documents like research articles, practitioner-oriented papers, case studies, transcribed interviews, white papers and technical reports (Coners and Matthies 2014; Indulska et al. 2012; King 2009; Freitas and Moscarola 1998). Analysing these publications leads to the recognition of a fast-growing collection of insights concerning various theories (Crane 2013; Chauvel and Despres 2002), schools of thought (Earl 2001), concepts or topics (Scholl et al. 2004). Due to the lack of taxonomy, diverse

terms are used synonymously to explain similar concepts (Nie et al. 2009). As a consequence, scholars have applied various research methods (O'Flaherty and Whalley 2004; Markus 1997) to achieve an improved understanding of KM. Among those, a popular method established as an empirically grounded method (Mayring 2010) is content analysis (CA), which fits well the analysis of huge data volumes and the handling of context-sensitive information (Krippendorf 2013).

CA can be applied manually (Heisig 2009; Bontis 2003) and in an automated way by using software solutions (Ribière and Walter 2013) – the latter providing advantages over the former like decreased analysis time and increased objectivity of the results (O'Flaherty and Whalley 2004). While automated CA and other text analytic methodologies have been continuously gaining importance over the past decades (Fisher et al. 2010), a problem concerns the use and analysis of compound concepts since automated CA typically yields frequency counts for single words only (e.g. *organisational learning* vs. *organisational* and *learning*). While the use of a dictionary (i.e. a list of relevant compound concepts like *organisational learning*) in general might solve this problem (Boritz et al. 2013; Ceci and Iubatti 2012; Vasalou et al. 2011; Gottschalk et al. 2005), there is paucity of research concerning the impact of dictionaries on CA. Thus, the goal of this paper is to develop and evaluate a dictionary to be used for automated CA. We focus on the KM domain, more specifically, the analysis of 614 research papers obtained from the highly ranked *Journal of Knowledge Management* (JoKM; cf. Serenko and Bontis (2009, 2013)). We perform CA using the automated approach in two ways – with and without using a dictionary. We thus aim to answer the following research question:

What are the benefits and pitfalls of using a dictionary for automated CA?

We compare results in terms of result similarity and result relevance. We also compare the dictionary-based approach to an analysis of the keywords of the publications considered. As a by-product, we extract core research themes within the *JoKM*. We thus provide an overview of key research topics within a premier outlet in the KM domain. To the best of our knowledge, this is the first documented attempt to investigate the effects of applying a self-developed dictionary to automated CA. We analyse the applicability of the combined approach and compare results to automated CA without using a dictionary when analysing large data sets in terms of time savings and quality of results. While researchers gain a basis for the systematic comparison and classification of research results, practitioners are supplied with an overview of current KM themes.

In Section 2, we provide an overview of CA, describe its application in previous research and highlight the research gap we address in our inquiry. We explain our research approach in Section 3, describing the development of the dictionary, the approach to automated CA as well as the analysis of the results. In Section 4, we present the results of the CA of papers from the *JoKM* differentiating between the results of automated analyses with and without a dictionary. Furthermore, we compare the results of the automated CA based on the dictionary to the keywords identified in the publications and describe common research themes. We discuss limitations of our study and implications of our research in Section 5. Our paper ends with a brief conclusion in Section 6.

2 An Overview of Content Analysis Method

Krippendorf (2013, p. 24) defines CA as a *"research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use"*. As an empirically grounded research technique, CA helps researchers in gaining new insights from available data and thus to obtain a deeper understanding of concrete phenomena. Moreover, this research method fits well the analysis of huge data volumes and the handling of context-sensitive information (Krippendorf 2013). According to Hopkins and King (2010) as well as Krippendorf (2013), the beginnings of CA in means of a systematically conducted text analysis can be traced back to the 17th century when the Catholic Church conducted text analyses during its inquisitorial pursuits to identify printed non-religious texts. Since then, CA has undergone several evolutions, which led to today's understanding

of CA as the analysis of any textual data – written or electronically provided via the World Wide Web (Dumay and Cai 2014; Krippendorf 2013).

Within the KM discipline, previous research has manually applied CA to investigate different research phenomena. In an attempt to harmonise KM frameworks, Heisig (2009) manually applied CA to compare and analyse 160 worldwide collected KM frameworks with regard to the following categories: reference (in the sense of title, author and year), country and region, type of the framework (descriptive, prescriptive or hybrid), knowledge definitions, frequently mentioned KM activities and critical success factors for KM. Data were coded and counted with regard to the occurrence frequency. The results were presented using graphs and tables. This procedure enables the comparison of the investigated frameworks, highlighting differences and correspondences between them. Moreover, Bontis (2003) examined intellectual capital (IC) disclosure by manually applying CA to annual reports of 10,000 Canadian corporations. For the purpose of the study, an IC researcher panel created a list of IC-related terms. The reports were searched with regard to the occurrence frequency of these terms. A table visualised the terms along with the number of Canadian corporations mentioning these terms in their reports. Based on the results, Bontis (2003) formulated several recommendations to help the corporations focus their efforts on augmenting IC disclosure through strategic and tactical initiatives.

While the findings of aforementioned studies yield helpful guidance, such manual analyses are time-consuming and can cause high failure rates due to human faults (e.g. absence, over-sighting and relevance-misestimating). These influences can significantly affect the validity of the results (Indulska et al. 2012; O'Flaherty and Whalley 2004). Furthermore, the subjectivity, experience and knowledge degree of the human coder play a particular role when conducting the analysis. Consequently, manual CA should be performed independently by at least two humans yielding adequate inter-coder reliability (Miles and Huberman 1994). The emergence of automated approaches and software to support analyses of almost all kinds and extent of data offers new opportunities to move from manual analyses to automatically supported analysis processes (Richards 2002). As highlighted by King (2009), the absence of studies exploring the use of automated CA in the IS discipline motivates the need to shed light on this research topic.

A first example of using automated CA is the study by Ribière and Walter (2013), who conducted CA on all 235 articles published in *Knowledge Management Research and Practice* Journal between 2003 and 2012. They focus on concepts, themes and keywords. While the keyword analysis was conducted by self-reporting the top 40 keywords (which occur in at least three articles) along with their frequency counts, text analysis was performed using the tool *Leximancer*. Using machine-learning methods, the authors applied the tool to visualise top occurring concepts, their frequency counts and relationships in form of a concept cloud. Semantically linked concepts built different clusters of themes, which were represented in a concept map. Utilising this research methodology allowed to identify the coverage of KM-related topics by the Journal *Knowledge Management Research and Practice* as well as KM trends within the observed time period (Ribière and Walter 2013).

Reflecting the evolution of CA, O'Flaherty and Whalley (2004) discuss the benefits and pitfalls of using software for CA. One of the main advantages is the ability of software to handle data albeit their type and volume. Subject to few basic transformations, all kind of structured or unstructured data like interviews, audio files, observation materials and journal publications can be analysed. The benefits of using software include reduced printed materials, decreased time required to manage the data and increased objectivity of the results. Consequently, analysts can spend more effort on the analysis process itself and have more opportunities to experiment with different analytic approaches and technological functionalities. A problem concerns the use and analysis of compound concepts within the automated analysis (e.g. while an automated search in general yields frequency counts for the terms *organisational* and *learning*, using a dictionary (i.e. a list of relevant compound concepts) also yields the frequency counts for compound concepts like *organisational learning*). Applying a predefined dictionary helps to handle this problem (Boritz et al. 2013; Ceci and Iubatti 2012; Vasalou et al. 2011; Gottschalk et al. 2005). For instance, Kaufmann and Bathen (2014) present such a dictionary-

based approach to identify topics and trends in the IS discipline using big data technologies. Relying on a dictionary-based approach to CA (also called 'bag-of-words' model) refers to a mapping algorithm that compares texts to words, phrases or sentences in a dictionary (Li 2010; Manning and Schutze 1999). Boritz et al. (2013) exposed the benefits of using means like dictionaries for automated CA. Dictionaries allow automated CA of unstructured text. While the use of dictionaries requires upfront cost for its development, additional documents can be analysed with little additional effort. Further benefits of applying dictionaries for automated CA include transparency and replicability of analyses. In contrast, manual analyses typically rely on explicit search rules, which are unlikely to be complete. Despite the widely automated approach, coding can be unreliable if no one with expertise reviews linkages. Obviously, the effectiveness of the approach highly depends on the dictionary used, that is, the list of words that is used for the mapping of texts (Kearney and Liu 2014). Problems can occur regarding the restricted and limited material scope as well as the limitation of using a fixed number of compound concepts which are pre-defined in the dictionary by disregarding new concepts which could emerge later (Kruschke 1992; Hampton 1995; Indulska et al. 2012).

Previous research (Boritz et al. 2013; Ceci and Iubatti 2012; Vasalou et al. 2011; Gottschalk et al. 2005) has focussed on using dictionaries, rather than analysing the suitability of using dictionaries. While the use of automated approaches for CA is considered to have several advantages compared to manual analyses, there is lack of research concerning the effects of using a dictionary for automated CA.

3 Research Approach

In this section, we describe our three-step research approach. The first step is to develop a dictionary of KM-related compound concepts. We describe the development of a dictionary, that is, the identification of a list of relevant compound concepts in the KM domain. Second, we explain the application of automated CA to the 614 articles published in the *JoKM* between 2004 and 2013 with and without the dictionary. Here, we use the dictionary developed in the first step. Finally, we describe the comparison of the results concerning both approaches. An overview of our research approach is shown in Figure 1.

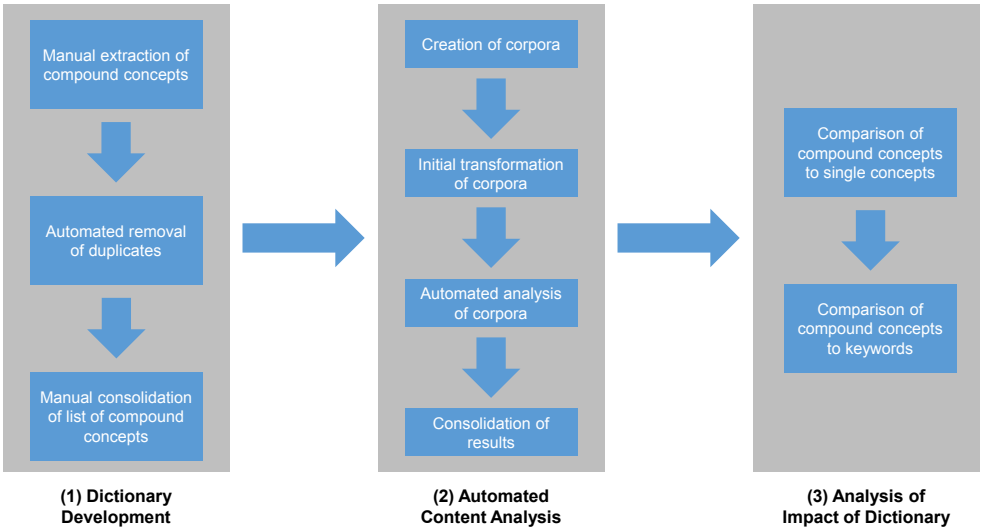

Figure 1. Three-step research approach.

3.1 Dictionary development

For the development of a dictionary (i.e. the list of compound concepts for the automated CA), we followed a manual approach that was performed by two research assistants. First, we reviewed titles, abstracts and keywords of the 614 published articles in the *JoKM* between 2004 and 2013. Within this

step, we identified compound concepts (e.g. *knowledge sharing*, *community of practice, organisational performance*), while neglecting words without references to the articles' content (e.g. *of*, *purpose*, *paper*, *abstract)*. Our primary focus was the extraction of nouns and nouns in combination with verbs (e.g. *transferring knowledge)* or adjectives (e.g. *organisational performance*) since such combinations provide relevant information concerning KM. We ignored pronouns, adverbs and prepositions (e.g. *them, they, very, in, of, and*) due to their low value of information (Baeza-Yates and Ribeiro-Neto 1999). After coding ten articles, one of the authors met with the research assistants to discuss preliminary results before coding all articles. Another refinement took place after the next 20 articles. We increased objectivity of the results by combining coding results from both research assistants.

Second, we straightened up the list of compound concepts by automatically removing duplicates, converting the concepts to lower cases and removing hyphens. While the former (i.e. removal of duplicates) served the purpose of shortening the list of compound concepts, the latter two transformations were necessary preparations for the automated CA (cf. Section 3.2). This data cleaning process resulted in a dictionary consisting of 3,847 concepts. The dictionary served as input for performing automated CA, which is described in the next section.

3.2 Automated content analysis

As basis for the automated CA, we used the same metadata as in the manual development of the dictionary (cf. Section 3.1), that is, publications' titles, keywords and abstracts. We assigned each publication a unique identifier to ensure consistency of the results when, for instance, querying the data. For an easy handling of the data, we created a .csv-file, containing all articles and selected metadata.

First, we built two different corpora. The first corpus *(Keyword_Corpus)* contains the publications' keywords only, whereas the second *(AbstractTitle_Corpus)* contains the publications' titles and abstracts. We chose to use these two different corpora for the following reasons. While titles and abstracts are presented in form of running text, keywords are separated by semicolons and thus require different handling of data. Another reason for handling data in different corpora is linked to the problem that occurs during the analysis of compound concepts like *organisational performance* or *intellectual capital*. While compound concepts in the *Keyword_Corpus* are identified by their separation through semicolons, software analyses on the *AbstractTitle_Corpus* is restricted to single words and neglects compound concepts.

Second and before performing automated CA, we transformed the corpora. The transformation includes the harmonization of all letters to lower case as well as removing punctuations and numbers within running text. These transformations are in line with the ones used in developing the dictionary (cf. Section 3.1) in order to perform a consistent analysis. This consistency is necessary since the case-sensitive automated CA does not recognise a term in a corpus, which does not exactly match a term in the dictionary (e.g. using the compound concept *Inter-organisational knowledge transfer* cannot be identified during text analysis if it is provided in the dictionary as *inter organisational knowledge transfer*). Due to hyphens and case sensitivity, we performed all transformations on both the dictionary and the corpora (e.g. transformations as letter harmonisation to lower case and removal of punctuation were applied to the dictionary and the corpora, thus resulting in, for example, a consistent spelling of concepts such as *inter organisational knowledge transfer*). Terms which occur in different spelling forms (singular/plural or British/American spelling) were included in the dictionary in all forms to be consolidated manually after performing the automated analysis.

Third, we performed automated CA by using the statistical computing software *R*, which is in an open-source programming environment for data analysis and visualisation (Venables et al. 2014) and provides various statistical packages for data and text analysis. Due to its platform independence and the large number of packages and features, *R* is among the most popular statistical computing software tools (Feinerer et al. 2008). We decided to use *R* based on a comparison of different tools (either through reading available research papers or through installing a free trial of the software) in terms of

available functionality and their adaptability and extensibility to fit the purposes of our research question. The crucial arguments to use *R* were that it provides a large set of functionality needed for our analysis and that it is possible to program necessary and not yet existing functionality on our own. Amongst others, *R* includes a text-mining package called *{tm}* that provides the functionality required to perform several transformation and analysis operations on textual data.

Our automated CA consisted of three different analyses. The first analysis was performed on the *Keyword_Corpus* and led to a list of all keywords along with their frequency count. The other two analyses were both performed on the *AbstractTitle_Corpus.* On the one hand, we applied automated CA using *R* without the dictionary. On the other hand, we applied automated CA using *R* with our dictionary as input. The dictionary approach works as follows. Each compound concept in the dictionary was matched to the text in the titles and abstracts. Each time the software identifies a match, the corresponding concept is stored. Finally, equal concepts were automatically accumulated resulting in a list which contains all concepts in conjunction with their frequency counts.

Finally, we aggregated the results of the dictionary-based approach by merging compound concepts. Semantically related concepts were consolidated manually (e.g. we merged *knowledge sharing* and *sharing of knowledge* to *knowledge sharing* and added their frequency counts). Additionally, we manually harmonised the occurrence of singular and plural forms of concepts to the singular forms (e.g. we merged *case study* and *case studies* to *case study* and added up their frequency counts) as well as different spelling forms of the concepts (e.g. due to British and American spelling). As a result, we retrieved a list of 2,723 concepts.

3.3 Comparison of results

Our comparison of the results primarily builds on the juxtaposition of the (compound) concepts that have been identified most frequently in our sample of articles published in the *JoKM* between 2004 and 2013. In order to show the impact of using a dictionary for automated CA, we make a two-fold differentiation. In the first step, we compare the results of single concepts and compound concepts, that is, the difference of analysing titles and abstracts with and without the application of a dictionary. We thus address the benefits of analysing publications in terms of compound concepts.

In the second step, we compare the most frequent compound concepts with the most frequent keywords of the publications. We thus analyse whether the compound concepts used in titles and abstracts equal those used in publications' keywords. With this comparison, we assess whether dictionaries can be built by simply extracting the keywords of all articles (rather than using the effort-intensive approach described in Section 3.1).

4 Content Analysis of Knowledge Management Research

In this section, we present the results of our analyses. First, we present a comparison of the results concerning our analysis of keywords and the analysis of titles and abstracts with and without using the developed dictionary. Second, we highlight common themes in *JoKM* over the past decade.

4.1 The impact of using a dictionary for automated CA

Table 1 shows the list of the top 30 identified concepts in the automated CA. The table is divided into three categories. The first two categories refer to our analyses of the *AbstractTitle_Corpus*. While the first category lists the top 30 concepts identified by using automated CA without considering the dictionary, the second category shows the top 30 concepts identified by using the dictionary for automated CA. The third category provides the top 30 keywords identified in the analysis of the *Keyword_Corpus*. For the concepts in each category, the frequency count (#) is provided. In the following, we describe both differences between compound and single concepts as well as the differences between compound concepts and keywords.

	AbstractTitle_Corpus				*Keyword_Corpus*	
	Without Dictionary		**With Dictionary**		**Keyword Analysis**	
Rank	**Concept**	**#**	**Compound Concept**	**#**	**Keyword**	**#**
1	Knowledge	4,917	Knowledge management	1,183	Studies	577
2	Management	1,530	Knowledge sharing	507	Knowledge management	559
3	Paper	1,215	Knowledge transfer	332	Organisation theory	96
4	Research	1,027	Case study	224	Information sharing	65
5	Study	885	Tacit knowledge	198	Organisational behaviour	62
6	Organisation	816	Community of practice	178	Innovations	60
7	Based	643	Knowledge creation	145	Organisational learning	54
8	Organisational	616	Knowledge management system	141	Corporate culture	34
9	Process	609	Information communication technology	137	Research & development (R&D)	32
10	Purpose	606	Organisational performance	110	Information technology	30
11	Sharing	597	Knowledge work	102	Knowledge	29
12	Approach	553	Social capital	94	Models	29
13	Finding	525	KM practice	89	United States (US)	28
14	Model	487	Decision making process	69	Competitive advantage	27
15	Value	479	Management system	67	Information systems	27
16	Can	450	Organisational knowledge	67	Intellectual capital	25
17	Implication	448	Organisational culture	67	Strategic planning	24
18	Practice	444	Business process	67	Strategic management	23
19	Design	424	Intellectual capital	66	Human resource management	21
20	Methodology	388	Management practices	63	Multinational corporations	21
21	Innovation	380	Knowledge city	58	Social network	21
22	Transfer	379	Knowledge flow	56	Product development	20
23	Framework	376	Literature review	55	Competition	19
24	Literature	337	Competitive advantage	47	Social capital	19
25	Development	336	Social network	41	Case study	18
26	Analysis	327	Dynamic capability	40	Decision making	18
27	Performance	322	Knowledge management process	39	Analysis	17
28	Originality	308	Organisational learning	39	Research	17
29	Social	290	Knowledge management initiative	39	Cities	16
30	Practical	280	Knowledge acquisition	38	Communication	16

Table 1. Top 30 identified (compound) concepts in applying automated CA (with and without using the developed dictionary) to the AbstractTitle_Corpus as well as performing automated CA to the Keyword_Corpus.

The concepts in the first category of Table 1 generally have a higher frequency count compared to those in categories 2 and 3. This effect can be attributed to the more general concepts in Category 1 (e.g. *knowledge* or *management* in Category 1 vs. *knowledge management* in Categories 2 and 3). Due to the lower frequency counts, the identification of trends (i.e. commonly used compound concepts) in

Category 2 may thus seem less likely. However, the quotient between the lowest and highest ranked concepts in Table 1 only marginally differ when comparing concepts and compound concepts (5.7% vs. 3.2%). Although the total frequency counts are generally lower, significant differences exist concerning the frequency counts when using a dictionary. It is thus also feasible to identify key research topics. On content level, we see the compound concepts to be more adequate to describe the KM research since the compound concepts add context to single concepts (*capital* vs. *intellectual capital*).

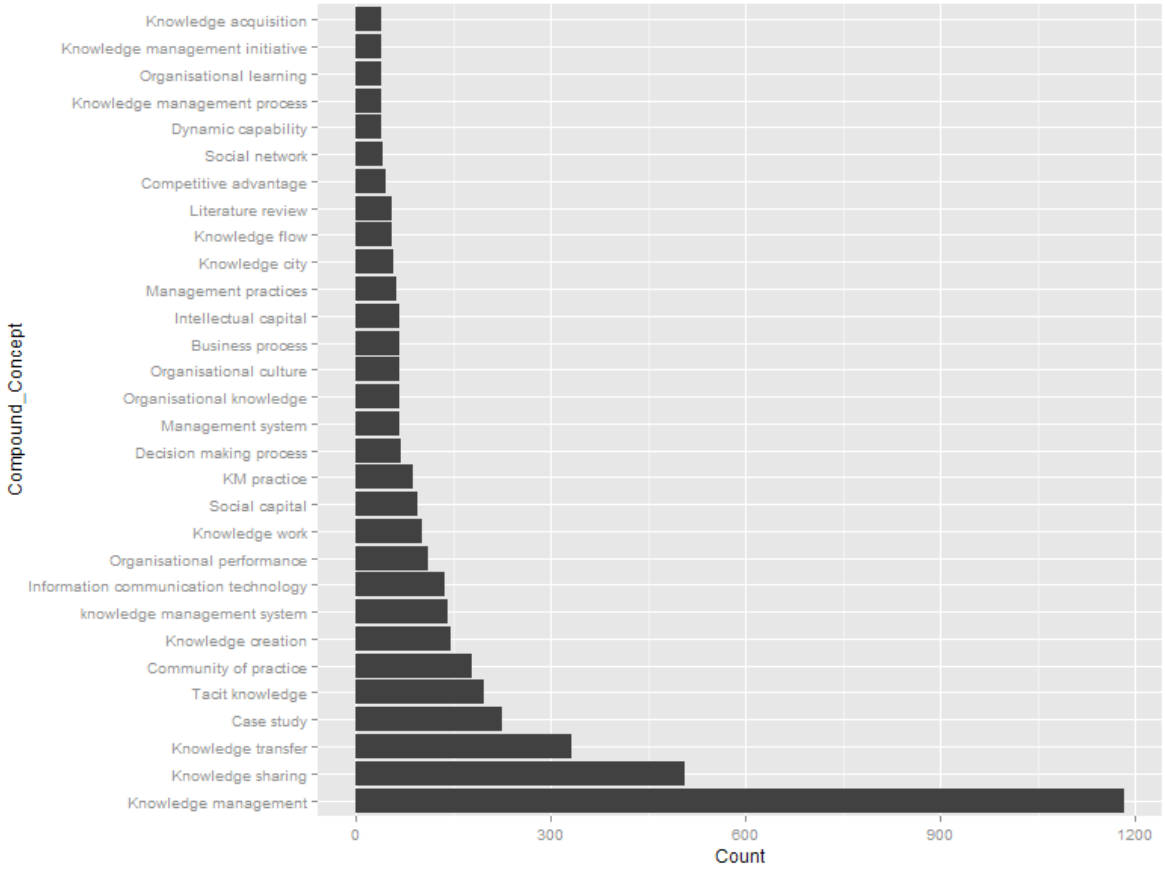

Figure 2. Distribution of the top 30 compound concepts with their frequency counts.

Considering the differences between compound concepts and keywords, the former occur more frequently. This can be explained by the fact that keywords only occur once per publication, whereas compound concepts can occur more often (e.g. once in the title and once in the abstract of the same publication). By comparing the frequency counts of Category 2 (compound concepts) with Category 3 (keywords), we observe differences regarding the ranked concepts. Many concepts (e.g. *knowledge transfer*, *knowledge sharing*, *communities of practices*, *knowledge management system*) are among the top 30 compound concepts without being listed in the top 30 keywords. We also recognise that frequently used keywords in publications (e.g. *organisation theory*) have not been used as compound concepts in titles or abstracts. In total, our analysis yielded 464 (compound) concepts in our dictionary that have not been identified as compound concepts in Category 2. These 464 (compound) concepts are keywords which occurred neither in the titles nor in the abstracts. Concluding, whereas also similarities exist between compound concepts and keywords used (e.g. *case study* is frequently applied in both ways), there are several major differences, which indicate that it is insufficient to generate dictionaries from the keywords used for the publications. For better comprehensibility, Figure 2 visualises the distribution of the top 30 compound concepts according to their frequency counts. Few

concepts can be seen as outliers in this context. *Knowledge management, knowledge sharing and knowledge transfer* have a significantly higher frequency count compared to the other 27 compound concepts (i.e. the added frequency count of the three concepts in relation to the frequency count of all 30 compound concepts in Category 2 amounts 46%).

Figure 3 shows the top 30 keywords and their occurrence frequency. The concept *knowledge management* is ranked under the first top 3 concepts with a probability of occurrence of 28% in relation to the overall probability of the top 30 keywords.

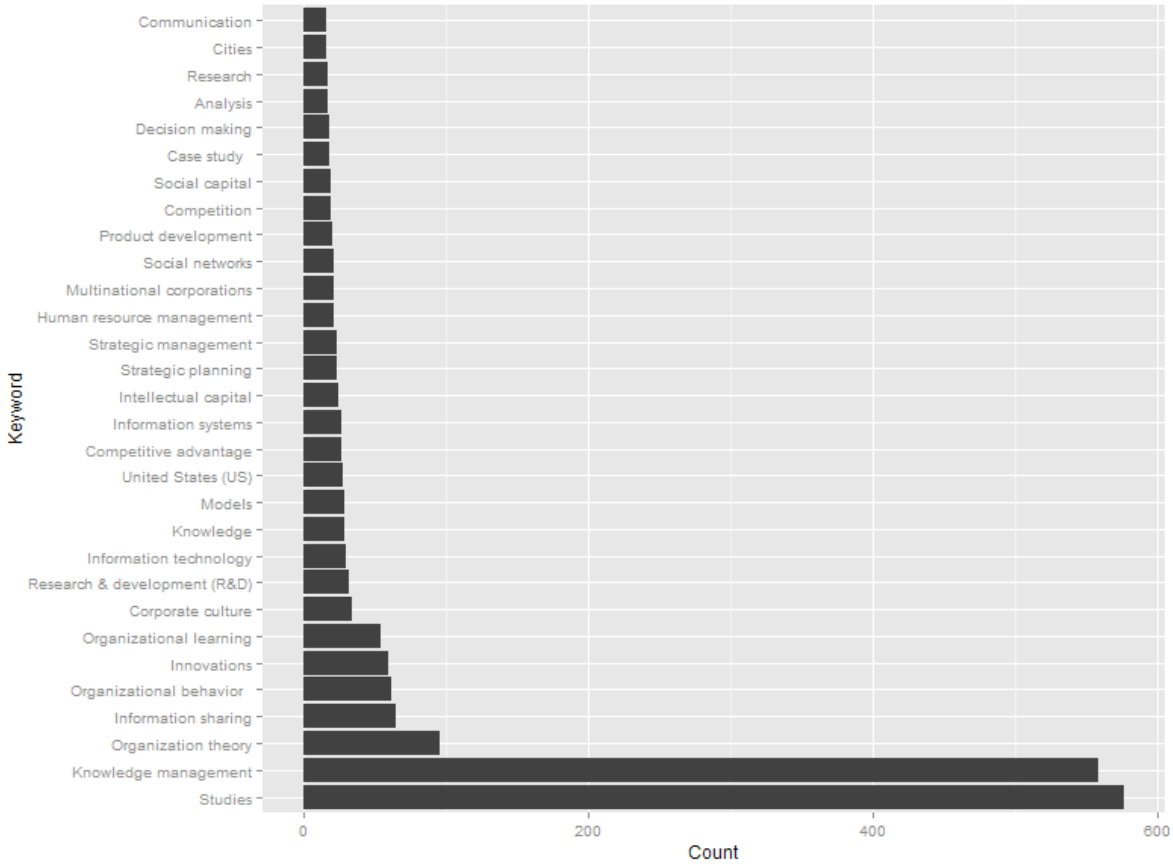

Figure 3. Distribution of the top 30 keywords with their frequency counts.

4.2 Key KM research topics

Taking a combined perspective to compound concepts and keywords, we were able to identify common themes to KM research in the *JoKM*. For this analysis, we omitted the concept *knowledge management,* which occurs with a much higher occurrence frequency than the remaining concepts (1. *knowledge management:* Count: 1183, 2. *knowledge sharing:* Count: 507) and thus causes biased results. Table 2 shows the ten most common concepts and their combined frequency count in titles, abstracts and keywords.

Rank	Concept	#
1	Knowledge sharing	507
2	Knowledge transfer	332
3	Case study	242
4	Tacit knowledge	198
5	Community of practice	178
6	Knowledge creation	145
7	Knowledge management system	141
8	Information communication technology	137
9	Social capital	113
10	Organisational performance	110

Table 2. Top 10 KM research concepts.

5 Discussion

Our study yields three major findings. First, with our analyses, we have shown that the use of a dictionary in automated CA is necessary to assess previous publications in a meaningful way. Second, we have shown the differences between keywords and compound concepts in titles and abstracts. Finally, we identified common concepts in KM research. In the following, we discuss limitations and directions for future research as well as implications of our study.

5.1 Limitations

In this section, we describe limitations of our study. More concretely, we refer to selection bias, multiple count bias, stop word bias and missing linkages of publications that enable identification of coherences among compound concepts and time series analysis.

A first limitation concerns the number of papers used in our study, in both the development of dictionary and performing the automated CA. We focussed on papers from one decade in a single journal only. Consequently, our findings have a limited generalizability. However, we carefully chose a highly ranked journal that can be seen as representative for the KM domain (i.e. *JoKM*; cf. Serenko and Bontis (2009, 2013)). While the journal to develop the dictionary and the journal to which we applied the dictionary-based CA were the same, our first suggestion for future research is to apply our dictionary to further KM journals in follow-up studies. Within this inquiry, it will be interesting to analyse in how far other journals extend the dictionary developed on the basis of JoKM. While our dictionary is based on the 614 publications in the *JoKM* between 2004 and 2013, its applicability to other journals is yet to be determined. However, since we have chosen a representative outlet and research themes are likely to occur in various outlets within a common domain, we are confident that the majority of the compound concepts in our dictionary are applicable to other KM journals as well.

Another limitation concerns the multiple count procedure of concepts within the *AbstractTitle_Corpus.* Concepts were counted as many times as they occur within the titles and abstracts. For instance, if the compound concept *knowledge sharing* occurred five times within a single abstract, it was counted five times. While future research should adapt our approach to automated CA, our results provide a first indication that compound concepts provide more context-specific insights (Category 2 in Table 1) compared to concepts (Category 1 in Table 1) and that keywords (Category 3 in Table 1) should not be solely used to develop dictionaries for automated CA, at least not in the context of KM research.

We further acknowledge that some (compound) concepts were expected to occur frequently within titles and abstracts of the analysed publications. For instance, it is not surprising that – in our sample of articles from the *JoKM* – the terms *knowledge* and *management* are the most frequent concepts

(Category 1) and that *knowledge management* is the most frequent compound concept (Category 2). This finding is in line with previous research studies (e.g. Heisig 2009, Ribière and Walter 2013). To avoid the influence of selected compound concepts on the overall distribution (e.g. the outliers in Figure 1), stop word lists (i.e. concepts that are omitted in automated CA) should be used when applying automated CA. The effect might be even stronger when using concepts (cf. Category 1 in Table 1). While we anticipated the influence of the compound concept *knowledge management*, one step to further develop the application of dictionaries in automated CA should be the development of a stop word list.

We point to the limitation of our analysis of key KM research topics (cf. Section 4.2). Considering our list of the top ten concepts in titles, abstracts and keywords (cf. Table 2) might lead to the assumption that research predominantly addresses case studies concerning the sharing and transfer of tacit knowledge in communities of practice with the support of information technology to increase social capital and organisational performance. However, our analysis does not account for interdependencies of the concepts. While we believe such an analysis to be worthwhile, it is beyond the scope of our analysis of the impact of dictionaries on automated CA.

5.2 Implications for future studies

Our study first provides several implications that should be taken into account when continuing research concerning automated CA.

First, we encourage researchers to use dictionaries when applying automated CA. Our results show that key research topics can be easily identified when using compound concepts in automated CA. Compound concepts that occur in titles and abstracts of the publications do not equal the keywords provided for the respective publications. We thus suggest that dictionaries should not be developed based on keywords only.

The comparison of the frequency counts in Category 2 (compound concepts) and Category 3 (keywords) reveals that several concepts (e.g. *knowledge transfer*, k*nowledge sharing*, c*ommunities of practices*, k*nowledge management system*) are among the top 30 compound concepts without being listed in the top 30 keywords. A possible explanation for this deviation is authors' attempt to choose common universal keywords to explain and cover concrete topics discussed in their papers. An example is the usage of the concept *information systems,* which may have been chosen as a universal concept instead of *knowledge management systems.* Another possible explanation could be the explicit suggestions of the *JoKM* to choose keywords, which might not appear in the paper's title and to use synonyms to ensure wider range of search terms[1]. We thus recommend that future studies address this issue, particularly when analysing publications that do not have such a suggestion.

We believe that the use of dictionaries is adequate to conduct meaningful CA on large sets of publications. In line with previous research by Boritz et al. (2013), we acknowledge that the development of dictionaries requires great effort but is an activity that typically needs to be performed only once. On the other hand, the automated approach provides benefits such as reduced error probability and time savings for the actual analysis. Moreover, our approach requires further effort to check semantic correspondences (e.g. *knowledge transfer* and *transferring knowledge*; singular vs. plural forms of compound concepts). Consequently, the higher objectivity of automated CA (O'Flaherty and Whalley 2004) is an aspect that is threatened by the manual consolidation. Nevertheless, it is unlikely that automated approaches can be conducted without human analysis (Boritz et al. 2013). Applying our dictionary to other KM outlets is a step for future research that is

[1] http://emeraldgrouppublishing.com/authors/guides/promote/optimize1.htm (Visited on 26/11/2014)

likely to reveal whether adaptions of the dictionary are necessary in order to apply the dictionary to other outlets.

We acknowledge that insights into topics of interest within the publications of *JoKM* between 2004 and 2013 are a by-product of our study. As such, our study does not provide in-depth insights as provided by other types of content analysis (e.g. Gable 2010). Primarily, this lack of insights results from the focus of our study. Nevertheless, a common theme seems to be *tacit knowledge sharing*. We believe that this topic will continue to prevail due to the complexity of this process. However, current trends on mobile applications, business intelligence and cloud computing (Oxford Economics 2011) are not among the top concepts in KM research yet. In general, it might be helpful if research and practice would be mutually informing. On the one hand, practitioners might benefit from the analysis of current research trends in the KM domain (here, an analysis of the interrelatedness of the compound concepts might be most beneficial; cf. the last paragraph in Section 5.1). On the other hand, practitioners can provide their major themes as compound concepts for the dictionary to be used for automated CA. Future research should build on the results of this study, taking them as a starting point for further analyses. By considering time-related data (e.g. which concepts are most likely to appear in publication data of year 2010) and repeating the analysis, research trends are likely to be identified by exploring whether specific topics appeared and/or disappeared within a specific time span.

6 Conclusion

Continuing previous research on performing automated CA in general and the suitability of using dictionaries in particular, we have compared different analyses of KM research using the example of publications within the *JoKM* between 2004 and 2013. With our study, we thus contribute to the research stream promoting the use of dictionaries for automated CA by analysing the effects of dictionaries of automated CA. On the one hand, our findings are in favour of using a dictionary for automated CA due to improved context understanding and time savings. On the other hand, we acknowledge that a subjective influence cannot be completely avoided in our process of CA and that the effort to initially develop the dictionary is significant. While paving the way for future research on the application of dictionaries in automated CA – in particular within the KM domain – our study's limitations also guide future research in optimising the approach chosen in this study. The negative facets notwithstanding, our study is in line with research promoting the role of dictionaries in improving the results of automated CA in striving to achieve a common understanding in today's fast evolving research domains. As Serenko (2013) stated, KM is a scientific discipline and its identity is formed through the interaction and the activities of its different internal and external stakeholders (i.e. KM scholars and practitioners) who have an image of the discipline in mind. Research in the field of CA leads to discover and manifest the identity of the KM discipline and provides its stakeholders with useful insights and information concerning future research efforts. While we acknowledge the limited insights concerning research themes in this study, our research contrasts the benefits and pitfalls of using dictionaries in automated CA and thus guides future research in pursuing the development of advanced CA approaches.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. England. Addison Wesley Longman Limited.

Bontis, N. (2003). Intellectual Capital Disclosure in Canadian Corporations. *Journal of Human Resource Costing & Accounting* 7(1), 9-20.

Boritz, J. E.; Hayes, L. and Lim, J.-H. (2013) A Content Analysis of Auditors' Reports on IT Internal Control Weaknesses: The Comparative Advantages of an Automated Approach to Control Weakness Identification. *International Journal of Accounting Information Systems* 14(2), 138-163.

Ceci, F. and Iubatti, D. (2012) Personal Relationships and Innovation Diffusion in SME Networks: A Content Analysis Approach. *Research Policy* 41(3), 565-579.

Chauvel, D. and Despres, C. (2002). A Review of Survey Research in Knowledge Management. *Journal of Knowledge Management* 6(3), 207-223.

Coners, A. and Matthies, B. (2014). A Content Analysis of Content Analysis in IS Research: Purposes, Data Sources, and Methodological Characters. In: *Proceedings of the Pacific Asia Conference on Information Systems*. Chengdu, June 24-28.

Crane, L. (2013). A New Taxonomy of Knowledge Management Theory: The Turn to Knowledge as Constituted in Social Action. *Journal of Knowledge Management* Practice 14(1), 1-20.

Dalkir, K. (2013). *Knowledge Management in Theory and Practice*. 2nd Edition. The MIT Press.

Dumay, J. and Cai, L. (2014). A Review and Critique of Content Analysis as a Methodology for Inquiring into IC Disclosure. Journal of Intellectual Capital 15(2), 264-290.

Dwivedi, Y. and Venkitachalam, K. (2009) Exploring Current State and Diffusion of Knowledge Management (KM) Research. In: *Proceedings of the Pacific Asia Conference on Information Systems*, Paper 104, Hyderabad, July 10-12.

Earl, M. (2001). Knowledge Management Strategies: Towards a Taxonomy. *Journal of Management Information Systems* 18 (1), 215-233.

Feinerer, I.; Hornik, K. and Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5), 1-54.

Fisher, I. E.; Garnsey, M. R.; Goel, S.; Tam, K. (2010) The Role of Text Analytics and Information Retrieval in the Accounting Domain. *Journal of emerging technologies in Accounting* 7, 1-24.

Freitas, H. and Moscarola, J. (1998). Content Analyzing Qualitative Data on Information Systems. In: Proceedings of the 4th Americas Conference on Information Systems. Baltimore, 941-943.

Gable, G. (2010) Strategic Information Systems Research: An Archival Analysis. *Journal of Strategic Information Systems* 19 (1), 3-16.

Gottschalk, L. A. and Bechtel, R. J. (2005) Computerized Content Analysis of Speech plus Speech Recognition in the Measurement of Neuropsychiatric Dimensions. *Computer Methods and Programs in Biomedicine* 77(1), 81-86.

Hampton, J. A. (1995). Testing the Prototype Theory of Concepts. *Journal of Memory and Language* 34(5), 686-708

Heisig, P. (2009). Harmonisation of Knowledge Management – Comparing 160 Frameworks around the Globe. *Journal of Knowledge Management* 13, 4-31

Holsapple, C. W. and Wu, J. (2008). In Search of a Missing Link. *Knowledge Management Research and Practice 6(31), 31-40*

Hopkins, D. and King, G. (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science* 54(1), 229-247.

Indulska, M.; Hovorka, D. S. and Recker, J. (2012). Quantitative Approaches to Content Analysis: Identifying Conceptual Drift across Publication Outlets. *European Journal of Information Systems* 21, 49-69.

Kearney, C. and Liu, S. (2014) Textual Sentiment in Finance: A Survey of Methods and Models. *International Review of Financial Analysis* 33, 171-185.

King, W. R. (2009). Text Analytics: Boon to Knowledge Management? *Information Systems Management* 26(1), 87.

Kaufmann, J. and Bathen, L. (2014). Themen und Trends in der Wirtschaftsinformatik – Eine Analyse unter Einsatz von Big-Data-Technologien. In: Proceedings of the Multikonferenz Wirtschaftsinformatik, 146-153.

Krippendorf, K. (2013). *Content Analysis: An Introduction to its Methodology*. 3d Edition. London. Sage Publications Inc.

Kruschke, J. K. (1992). Alcove: An Exemplar based Connectionist Model of Category Learning. *Psychological Review* 99(1), 22-44.

Li, F. (2010) The Information Content of Forward-looking Statements in Corporate Filings – A Naive Bayesian Machine Learning Algorithm Approach. *Journal of Accounting Research*, 48, 1049-1102.

Manning, C. and Schutze, H. (1999) Foundations of Statistical Natural Language Processing. The MIT Press.

Markus, M. L. (1997). The Qualitative Difference in Information Systems Research and Practice. In: *Proceedings of the IFIP TC8 WG 8.2. Information Systems and Qualitative Research*. Ed. by Lee, A., Liebenau, J. and Degross, J. I., Philadelphia, USA.

Mayring, P. (2010). "Qualitative Inhaltsanalyse". In *Handbuch Qualitative Forschung in der Psychologie.* Ed. by G. Mey and K. Mruck. VS Verlag für Sozialwissenschaften, 601-613.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative Data Analysis*. Thousand Oaks. Sage.

Nie, K., Ma, T. and Nakamori, Y. (2009). An Approach to Aid Understanding Rmerging Research Fields – the Case of Knowledge Management. *Systems Research and Behavioral Science* 26, 629-643.

O'Flaherty, B. and Whalley, J. (2004). Qualitative Analysis Software Applied to IS Research - Developing a Coding Strategy. In: *Proceedings of the European Conference on Information Systems*. Turku, June 14-16.

Oxford Economics (2011). Digital Megatrends 2015: The Role of Technology in the New Normal Market.

Richards, L. (2002). *Using NVivo in Qualitative Research*. 3'rd Edition. Qualitative Solutions and Research Publications.

Ribière, V. and Walter, C. (2013). 10 Years of KM Theory and Practices. *Knowledge Management Research and Practice* 11, 78-91.

Scholl, W., König, C., Meyer, B. and Heisig, P. (2004). The Future of Knowledge Management: An International Delphi Study. *Journal of Knowledge Management* 8(2), 19-35.

Serenko, A. (2013). Meta-analysis of Scientometric Research of Knowledge Management: Discovering the Identity of the Discipline. *Journal of Knowledge Management* 17(5), 773-812.

Serenko, A. and Bontis, N. (2009). Global Ranking of Knowledge Management and Intellectual Capital Academic Journals. *Journal of Knowledge Management* 13(1), 4-15.

Serenko, A., Bontis, N., Booker, L., Sadeddin, K. and Hardie, T. (2010). A Scientometric Analysis of Knowledge Management and Intellectual Capital Academic Literature (1994-2008). *Journal of Knowledge Management* 14(1), 3-23.

Serenko, A. and Bontis, N. (2013). Global Ranking of Knowledge Management and Intellectual Capital Academic Journals: 2013 Update. *Journal of Knowledge Management* 17(2), 307-326.

The Economist (2010) Data, Data Everywhere. A Special Report on Managing Information, pp. 1-20.

Vasalou, A.; Gill, A. J.; Mazanderani, F.; Papoutsi, C. and Joinson, A. (2011) Privacy Dictionary: A New Resource for the Automated Content Analysis of Privacy. *Journal of the American Society for Information Science and Technology* 62(11), 2095-2105.

Venables, W. N.; Smith, D. M. and the R core team (2014). *An Introduction to R*. URL: http://cran.r-project.org/doc/manuals/R-intro.pdf.

Vorakulpipat, C. and Rezgui, Y. (2008). An Evolutionary and Interpretive Perspective to Knowledge Management. *Journal of Knowledge Management* 12(3), 17-34.