2016 49th Hawaii International Conference on System Sciences

Comparing European and North American IS Research Using Concept Frequency Analysis: Methods and Preliminary Findings

Daniel S. Soper and Ofir Turel
Information Systems and Decision Sciences Department
California State University, Fullerton
dsoper@fullerton.edu oturel@fullerton.edu

Abstract

This study introduces a computational text-mining technique called Concept Frequency Analysis, and demonstrates its usefulness by detecting convergence and divergence trends between North American and European information systems (IS) research. Applying this technique to the corpora of articles published in leading North American and European IS journals (MISQ and ISR vs. EJIS and JIT, respectively) from 1991 through 2013 (2,959 articles), using 10.6 million unique concept labels identified from Wikipedia, and performing approximately 31.4 billion document search operations, we find that while the conceptual density of research affiliated with European and North American research seems to be gradually converging, differences in the concepts and topics being discussed on these two continents have been growing since the mid-1990s. Hence, the world of IS research is not yet flat, and in fact appears to be shifting away from cross-continental convergence.

1. Introduction

In the mid-1920s, a great schism began to divide the world of theoretical physics. Whereas Albert Einstein and his acolytes held to the classical notion of a deterministic universe, other leading physicists, led principally by Niels Bohr and Werner Heisenberg, were beginning to formulate a new view of reality which was fundamentally rooted in the concepts of uncertainty and unpredictability, later to be known as the Copenhagen Interpretation [1]. Interested scholars of the era followed the development of these two divergent perspectives by reading the competing bodies of scientific literature published by each faction, and by digesting and debating the concepts and ideas advanced therein. Due to the comparatively small size of these two corpora of scientific literature, such scholars could readily identify both the common and the distinguishing concepts which characterized each school of thought. What would happen, however, if the size of these corpora were so large that the constraints imposed by time and human cognition would preclude the possibility of reading and digesting all of the scientific literature that they contained? How, then, might interested scholars reasonably hope to document, quantify, and ultimately understand the ways in which such large bodies of literature were different or alike?

In the modern era, the sheer volume of scientific research that is produced annually by scholars in the information systems (IS) field has made it impossible for a lone human being to read and fully digest every research article being published. Unfortunately, the reality of this situation means that without the aid of innovative computational tools, comparing two large bodies of text – such as the research published in leading European and North American IS journals – is simply not feasible without resorting to guesswork and speculation. For this reason, developing computational text mining tools with the capacity to analyze, objectively compare, and extract insights from large corpora of text is both timely and desirable.

In this paper we aim to (1) develop and introduce a computational text mining technique which we call *concept frequency analysis*, and (2) demonstrate the technique's power and usefulness by applying it to an interesting research question; to wit: Is the world of IS research, as expressed through concepts mentioned in journals headquartered in and led by people from different continents, flat? More specifically, we propose to examine possible differences over time between IS research published in North America and Europe. While the world seems to be converging to a global culture and is argued to be growing "flatter", and while there seems to be cross-continental pollination in IS research, evidence still remains regarding cultural differences among countries and continents [2-4]. As an example, in developing the Senior Scholar's Basket of 8 IS journals (http://aisnet.org/general/custom.asp?page=SeniorScholarBasket), the scholars recognize geographical diversity, and include both predominantly European (e.g., European Journal of Information Systems [EJIS], Journal of Information Technology [JIT]) and predominantly North American (MIS Quarterly [MISQ], Information System Research [ISR]) researn journals. Hence, it is interesting to consider the state of IS research in terms of

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.146

its cross-continental convergence or divergence over time.

Although there are many ways in which a journal can be geographically classified (e.g., modal affiliation of authors, their nationalities, modal affiliations of editors, their nationalities, location of headquarters / hosting university, affiliation of the editor-in-chief), we chose, for reasons of simplicity, to rely on two criteria. In our study, a journal is seen as primarily affiliated with a continent (Europe or North America), if it is headquartered on that continent, and if the editor-in-chief's professional affiliation is most closely linked with that continent. Using these criteria, we determined MISQ and ISR to be predominantly North American journals, and EJIS and JIT to be primarily European journals. Per our classification criteria, this does not mean that North American researchers cannot publish in European journals and vice-versa. It simply means that based on the two criteria described above, each journal is believed to be more strongly associated with one continent and its research culture than with the other.

2. On concepts

Prior to proceeding with the development of our concept frequency analysis methodology, it is first important to define what we mean by the term *concept*. There are at present three dominant theories which attempt to address and define the nature of concepts. The oldest and perhaps most prominent of these theories is the empiricist (or classical) theory of concepts, which has its roots in Aristotelian philosophy [5]. This theory holds that each unique concept can be defined in terms of a set of features, with each feature being both necessary and sufficient in order for a given entity to fall under the auspices of that concept [6]. For example, the concept of a bicycle might be defined as a human-powered, pedal-driven land vehicle with two inline wheels. According to the classical theory of concepts, an entity could thus be considered a bicycle if and only if it possessed all of these features. By extension, any entity not possessing all of these features could not be considered a bicycle.

Although the classical theory held sway for more than two millennia, arguments against the theory's propositions have in recent decades begun to erode its prominence. For example, imagine that the front wheel of a bicycle is removed in order to facilitate storage, transport, or repair. With its front wheel missing, is the entity still a bicycle? The classical theory of concepts would, of course, assert that because the entity does not have two wheels, it is no longer a bicycle; i.e., by removing the front wheel, we have effectively destroyed the entity's "bicycleness".

Issues such as these with the tenets of the classical theory led to the development of prototype theory, which, like the classical theory, holds that each unique concept can be defined in terms of a set of features, but that the concept will *tend* to possess each of those features, rather than being *required* to possess each of those features [7]. The boundaries of a concept class are hence fuzzy, and a given entity's membership in a concept class is determined by comparing its features to a prototypical reference class for the concept. Thus, a bicycle with a missing front wheel is, according to prototype theory, still a bicycle.

More recent work on the nature and structure of concepts has led to the development of theory-theory [6]. In the context of concepts, theory-theory holds that the boundary conditions for what constitutes a concept emerge from a process of internal theorizing [7]. The nature and structure of a concept are hence not only inextricably linked to the relationships that interconnect the concept to other concepts, but are also subject to a continual process of modification and refinement. A person might, for example, erroneously believe that a tomato is a vegetable, and hence partially define her concept of what constitutes a tomato by its relationship to "vegetableness". Upon learning not only that a tomato is actually a fruit, but also the reasons why a tomato is a fruit, our subject could be expected to refine and ostensibly improve her internal theories about both the concept of a tomato and the concept of fruit. From the perspective of theory-theory, then, a concept can be conceptualized as an entity about which theorizing can occur.

Theorizing about the nature of a concept does not occur in isolation, but is instead intimately interwoven with the relationships that connect one concept to another. In the context of our previous example, one could not, according to theory-theory, fully understand the nature and structure of a bicycle without also understanding the concept of a wheel, the concept of a pedal, and so forth. The view of concepts used in the current research project is based on the perspective put forth by theory-theory; to wit, that a concept is an entity about which theorizing can occur, and that a concept is defined in terms of its relationships to other concepts.

Inasmuch as our study involves a great deal of text analysis, a few additional considerations merit some attention. First, from a linguistic perspective, human beings have a propensity to assign multiple labels (i.e., coreferences) to a single concept. For example, *U.S.*, *USA*, *America*, *United States*, and *United States of America* are all textual coreferences that might be used in the English language to refer to the same underlying concept. Concepts must therefore be conceptualized as abstract entities to which one or more textual labels might apply, and text analysis of concepts must attend to this situation by means of coreference resolution [8].

Second, no universal agreement exists on the correct spelling of many English words, with thousands of notable differences existing between and within the written forms of British English and American English.

For example, *organizational behavior* (American English) and *organisational behaviour* (British English) are two textual labels that refer to the same underlying concept. Together, this multiplicity of textual labels presents a serious challenge if one hopes to accurately quantify the frequency with which any given concept appears in a large corpus of text. In the following section we address these and many other issues as we develop and present our analytical methods.

3. Methodology

One of the central theoretical tenants in linguistic semantics is that the nature of a written document can be characterized by the concepts that the document contains and the frequency with which those concepts appear [9]. Focusing on concept frequencies as a means of examining large corpora of text is a new approach to computational text analysis, however, and as such there are no conventional or universally accepted methods for performing such analyses. In this section we will therefore develop several methodological approaches for conducting concept frequency analyses on large textual corpora. To do so, we will rely on and extend the well-established methods which have been developed in support of a related, but more primitive form of text mining known as n-gram analysis.

3.1 Summary of n-gram analysis

In brief, an *n*-gram is a sequence of words of length *n* that is extracted from a larger sequence of words [10]. In an n-gram analysis, a software program examines a corpus of text and constructs a database containing all of the possible word sequences whose lengths are less than or equal to a pre-established maximum number of words. If, for example, the maximum allowable length of an n-gram were three words, then the sentence "Einstein won a Nobel Prize" would yield three 3-grams (*Einstein won a*, *won a Nobel*, and *a Nobel Prize*), four 2-grams (*Einstein won*, *won a*, *a Nobel*, and *Nobel Prize*), and five 1-grams (*Einstein*, *won*, *a*, *Nobel*, and *Prize*). By identifying all of the possible n-grams in a corpus of text, quantifying how often each n-gram appears, and recording the date (e.g., year) of the source document for each n-gram frequency record, it becomes possible to analyze trends in the usage of particular words or phrases over time [11, 12]. N-grams have been used as the basis of text analytic research in many fields, including the information systems field [13-15].

Unfortunately, n-gram analysis has many shortcomings which greatly constrain its usefulness. Notably, most of the n-grams that result from the n-gram construction process have little, if any, practical value. Of the 12 n-grams identified in the simple example above, it could be reasonably argued that only two (*Einstein* and *Nobel Prize*) qualify as concepts (i.e., entities about which theorizing can occur). Further, the imposition of a maximum n-gram length precludes the consideration of concepts with longer textual labels. The common maximum n-gram length of five words [12], for example, would not allow for the study of a concept such as "unified theory of acceptance and use of technology".

Finally, simple n-gram analysis does not account for co-references when computing n-gram frequencies. This means that if a researcher wanted, for example, to gain accurate insights into the frequency with which the European Union was mentioned in an a body of text, she would need to manually combine the frequencies for both "European Union" and "EU". In order to resolve these problems, we describe in the following subsections a text-analytic method which we call *concept frequency analysis*, and show how it can be usefully applied to compare European and North American IS research.

3.2 Overview of concept frequency analysis

As noted previously, the corpora of European and North American IS research have both grown to be so large over the past few decades that it is no longer feasible for a human being to be fully familiar with either of them. Without the aid of additional tools, any conclusions drawn by a human being regarding the similarities or differences between these two large bodies of scientific literature must therefore be speculative and hence unreliable. A very useful and computationally feasible approach to reconciling this problem is through the use of what we call *concept frequency analysis*.

Broadly speaking, concept frequency analysis seeks to identify and illuminate substantive similarities and differences among two large corpora of text by considering the concepts that they contain, as well as the frequency with which those concepts appear. From a theoretical perspective, two corpora are highly similar if the concepts that they contain largely overlap, while the corpora are highly dissimilar if the concepts that they contain overlap only slightly. This is illustrated using a set of Venn diagrams in Figure 1 below.

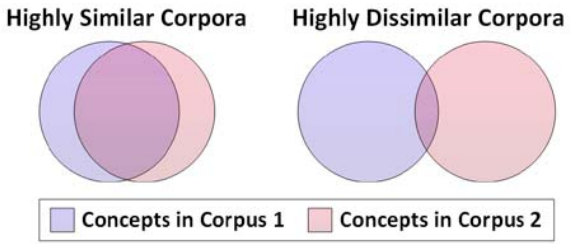

Figure 1.The relationship between concept overlap and similarity in two corpora of text.

Identifying similarities and differences in concept usage thus lies at the core of concept frequency analysis. Although the figure above depicts only a cross-sectional comparison of two corpora, concept frequency analysis

becomes truly powerful when extended into the realm of time series analysis (e.g., by studying the conceptual evolution of a corpus from year-to-year, or by comparing the conceptual nature of two corpora over time). Before performing such analyses, however, it is first necessary to build the corpora of text, construct a database of concepts, and compute the frequencies with which each concept appears in each corpus.

3.3 Building the corpora of text

Given that our analytic focus in the current project was on the degree of similarity between European and North American IS research over time, it was necessary to construct a large corpus of European IS literature and a large corpus of North American IS literature which could subsequently be compared with one another. For this purpose, we assembled an electronic collection of every research article that had been published in MISQ, ISR, EJIS, and JIT between 1991 and 2013. Articles from MISQ and ISR formed the North American corpus, while articles from EJIS and JIT formed the European corpus. These four journals were chosen for the analysis both because of their lengthy publication histories and because they are generally considered to be among the finest scholarly journals in North American and European IS research [16, 17]. At the same time, focusing on all journals in the senior scholars' basket of 8 journals would be prohibitive with the techniques we describe below. With respect to the timeframe used in the analysis, 1991 was chosen as the first year of the analytic timeframe because it was the first year in which all four of the journals were concurrently publishing research, while 2013 was used as the last year of the analytic timeframe because it was the last year for which complete data were available at the time when the corpora were constructed. In total, our collection of European and North American IS literature included 2,959 research articles spanning a 23-year publication history. Of the 2,959 total research articles, the North American corpus contained 1,306 articles, while the European corpus contained 1,653 articles.

After having assembled our electronic library of IS research articles, we next converted each article into a machine-readable format using the Adobe optical character recognition (OCR) algorithm, after which we were able to extract the complete text of each article. Excepting for acronyms, all of the words in each article were converted to lowercase so as to eliminate any problems that might otherwise arise due to capitalization.

3.4 Constructing a database of concepts

The next step in identifying and quantifying the concepts which appeared in the corpora of European and North American IS research was to construct a database containing a very large number of concepts for which to search during the text analysis process. For this purpose we began by downloading the complete set of article titles contained in the English language Wikipedia [18]. In light of the vast scope of this online encyclopedia, we reasoned that nearly every concept of even moderate importance would be likely to have an associated article in the English language Wikipedia. Although we acknowledge that Wikipedia does not contain an article for *every* concept, it nevertheless represents the largest collection of human knowledge ever assembled [19], and can therefore reasonably be expected to contain information about at least a sizeable proportion of all known concepts. At the time of our analysis, the English language Wikipedia contained 4,699,635 ordinary content articles.

As noted in the section discussing concept theory, each unique concept might have many different textual labels (e.g., the labels "HICSS" and "Hawaii International Conference on System Sciences" refer to the same underlying concept), and for this reason concepts in the database were modeled as abstract entities to which many different labels could be assigned. In addition to the ordinary content articles, Wikipedia also contained a large number of so-called "redirect" pages. These redirect pages serve as alternative names for ordinary content articles, and were hence used as alternate textual labels for the set of 4.7 million concepts. Further, in light of the many variations in spelling that exist between American and British English, it was necessary to construct a set of additional alternate labels for the concepts in the database which took these variants into account. Using the *Word List of US-UK Spelling Variants* [20], we therefore computationally constructed all possible alternate spellings for the concepts in the database, and added those alternate spellings as additional textual labels for each concept as appropriate. After completing these activities, the final concept database contained approximately 4.7 million unique concepts and 10.6 million unique concept labels.

3.5 Computing concept frequencies

After having completed the construction of the concepts database and the European and North American IS research corpora, we next searched for each concept label within the complete text of each research article, counting the frequency with which each label appeared as the process unfolded. For this purpose, we used a search strategy in which the concept labels were iteratively considered beginning with the textually longest labels and working toward the textually shortest labels. After counting the frequency with which each concept label appeared in an article, all instances of that concept label were removed from the article text, after which the next concept label would be considered. By proceeding in this manner, we were able to eliminate any problems associated with one concept label containing

the name of another concept label (e.g., the string "information systems theory" contains the substring "systems theory" – these are, of course, two very different concepts!). With 2,959 IS research articles and 10.6 million unique concept labels, a total of approximately 31.4 billion document search operations were necessary in order to fully scan the corpora. Upon completing the entire search process, the low-level concept label frequencies were appropriately aggregated into article-level concept frequencies, which were thence aggregated into yearly European and North American concept frequencies in support of the analyses described below.

3.6 Analyzing conceptual density

One of the interesting ways in which concept frequency analysis can be applied is to examine the *conceptual density* of a corpus over time. In the context of research articles, an article's conceptual density is computed as the number of unique concepts appearing in the article divided by the article's length (i.e., number of words). From an interpretive perspective, *conceptually dense* articles discuss or mention many unique concepts relative to their length, while *conceptually sparse* articles mention or discuss few unique concepts relative to their length. Since the average conceptual density of the text within a corpus can be readily modeled as a linear function of time, a standard moderation (interaction) analysis can be used to compare and study the conceptual density trajectories of two corpora of text over time. This approach was thus used to gain comparative insights into the nature of the writing which appears in European vs. North American IS research.

3.7 Analyzing conceptual convergence and divergence

The similarity (or difference) between two corpora of text can be studied by examining the extent to which the usage (i.e., the relative frequencies) of concepts differs in one corpus vs. the other corpus over time. In effect, such an analysis allows for a determination to be made regarding the extent to which the two corpora have addressed the same content over time.

To perform this examination, a two-sample, unpaired t-test was conducted wherein the average frequency with which each concept appeared in European IS research articles during a particular year was compared against the average frequency with which the same concept appeared in North American IS research articles during the same year. As an illustrative example, imagine that during a particular year 100 articles were published in leading European IS journals while 80 articles were published in leading North American IS journals. If a given concept appeared an average of 3.0 times per article with a variance of 1.0 appearances in the European journals while the same concept appeared an average of 2.5 times per article with a variance of 1.2 appearances in the North American journals, then a two-sample unpaired t-test would yield a p-value of 0.004, thus indicating that the frequency with which the concept appeared differed significantly between the leading European and North American IS journals during that year.

Due to the large number of hypothesis tests, the observed p-values were corrected using the Bonferroni method, which is generally considered to be the most conservative approach to controlling the familywise error rate [21, 22]. Bonferroni corrections were thus applied according to the number of statistical tests conducted for each year of the analysis.

4. Preliminary findings

In the following subsections we rely upon the methods described above to analyze not only the conceptual density of articles appearing in leading European and North American IS journals between 1991 and 2013, but also to quantify the extent to which the conceptual nature of IS research on the two continents has converged or diverged over time.

4.1 Conceptual density of European vs. North American IS research

Recalling our definition of conceptual density as the ratio of the number of unique concepts appearing in an article to the article's length (i.e., its number of words), we begin the presentation of our findings with an interaction analysis which directly compares the conceptual density of European and North American IS research between 1991 and 2013. The overall model was highly significant ($F_{3,2955} = 290.11$, $p < 0.001$, $R^2 = 0.23$), as was the interaction term ($p < 0.01$), thus revealing a statistically significant difference in the conceptual density of European and North American IS research over time. The specific nature of the conceptual density of the IS research on each continent is illustrated in Figure 2.

As shown in the figure, the conceptual density of European IS research has been substantially greater than the conceptual density of North American IS research over time, indicating that on average, articles published in leading European IS journals have addressed more unique concepts per article relative to their length than have articles published in leading North American IS journals. Further, the figure also indicates that the conceptual density of both European and North American IS research has been declining steadily over time, with both of these rates of decline being highly statistically significant ($p < 0.001$). The conceptual density of European IS research, however, has been declining at a significantly faster rate than that of North

American IS research, with the conceptual density of European IS research declining by approximately 36.5% during the 23-year span of our analysis, and North American IS research declining by approximately 31.4%. These trends point to a future convergence between Europe and North America with respect to the conceptual density of the IS research published in their leading journals.

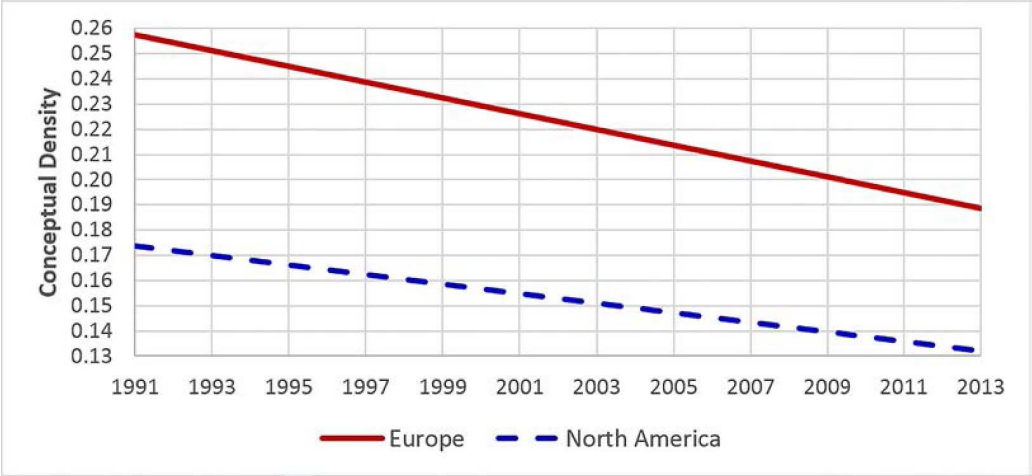

Figure 2. Conceptual density of European and North American IS research from 1991 to 2013.

Further insights into the above-described phenomena can be derived by examining the elements of which the conceptual density measure is composed; i.e., the number of unique concepts and the number of words in each research article. Two further interaction analyses were thus duly conducted which respectively examined (1) the number of unique concepts appearing in European and North American IS research between 1991 and 2013, and (2) the number of words appearing in each continent's IS research articles during the same timeframe. The overall models for both analyses proved to be highly significant ($p < 0.001$ in both cases), as did each model's interaction term ($p < 0.001$ for the 'number of concepts' model, and $p < 0.05$ for the 'number of words' model). The results thus reveal statistically significant differences between Europe and North America with respect to both the number of unique concepts and the number of words which have appeared in their respective IS research articles over time. These trends are illustrated in Figure 3 below.

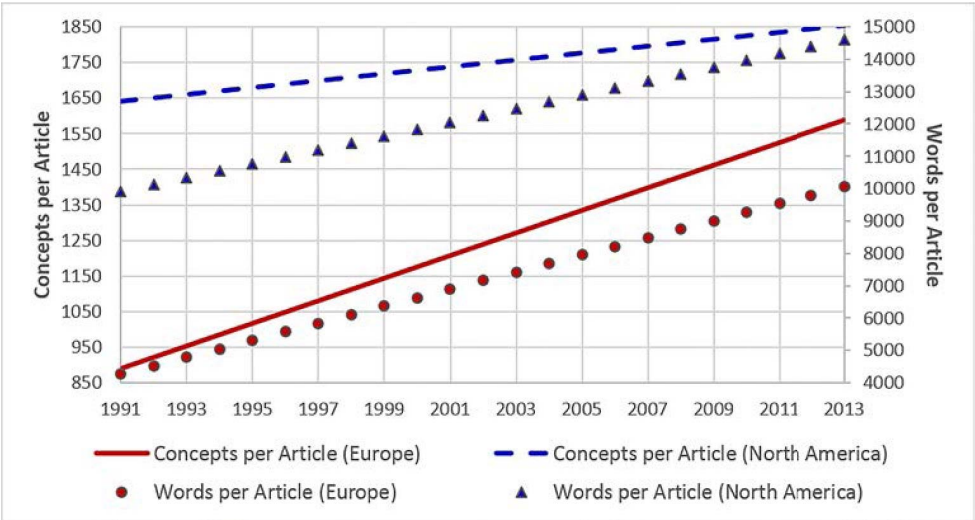

Figure 3. Concepts and words per article in European and North American IS research from 1991 to 2013.

In absolute terms, the figure above reveals that the average number of unique concepts per article has been growing for both European and North American IS research, with the rate of conceptual growth of

European IS research substantially outpacing that of North American IS research. Specifically, the average number of unique concepts in European IS research articles has been growing by 31.69 concepts per article per year, while by contrast the average number of unique concepts in North American IS research articles has been growing by 9.68 concepts per article per year. Comparatively speaking, this implies that European IS research has been embracing new concepts at a much faster rate than North American IS research. Put differently, the scope and breadth of topics addressed in European IS research have been growing much more quickly than have the scope and breadth of topics addressed in North American IS research. Nevertheless, in absolute terms the data indicate that the average number of unique concepts appearing in North American IS research articles continues to exceed that of European IS research articles. When considered together, this conceptual growth may be an indicator of the increasing importance and variety of IS in society, the increasing scope and complexity of the phenomena being addressed by IS research, or both.

As with the number of concepts per article, the average number of words per article has, in absolute terms, also been growing for both European and North American IS research. Specifically, the average number of words in European IS research articles has grown by 263.79 words per article per year, while the average number of words in North American IS research articles has been growing by 213.47 words per article per year. Since the average number of words in both European and North American IS research articles has been growing substantially faster than the average number of unique concepts per article, we can easily understand the observed decline in conceptual density over time presented at the outset of this subsection. To wit, although IS researchers have, on average, been discussing more and more unique concepts in their manuscripts over time, the length of those manuscripts has been growing far faster than the additional number of unique concepts, hence causing an overall decline in the conceptual density of IS research.

The results obtained from our analyses thus indicate that since the early 1990s, European IS research articles have, on average, contained substantially fewer words and unique concepts per article than have North American IS research articles. Both of these gaps are, however, rapidly narrowing, as the rates of growth in the number of unique concepts per article and the length of each article in European IS research both substantially exceed the analogous rates of growth in North American IS research. From an interpretive perspective, we do not view these phenomena as European IS research "catching up" to North American IS research, but rather as growing agreement among European and North American IS researchers – or at least among editors, reviewers and authors of leading IS journals – regarding the conceptual scope and complexity that should characterize IS research of the highest quality.

From a theoretical perspective, one might reasonably conjecture that as more and more concepts are included in a scientific manuscript, the author of the manuscript will, on average, require substantially more writing in order to satisfactorily integrate and discuss the additional concepts. Put another way, the data suggest that as the number of concepts in a scientific manuscript grows, the "overhead" writing costs incurred by the author can also be expected to grow, but at a disproportionally faster rate. Put yet another way, as the number of unique concepts appearing in a scientific manuscript grows, the conceptual density of the manuscript can be expected to decline. This is, of course, precisely what was observed above in our analyses of both European and North American IS research.

4.2 The convergence and divergence of European and North American IS research

As described previously, the similarity (or difference) between European and North American IS research can be studied in the aggregate by examining the extent to which the usage (i.e., the relative frequencies) of concepts differs in the IS research published in leading journals on these two continents over time. In effect, this allows for insights to be gleaned regarding the extent to which IS research published in leading European and North American journals has addressed the same content over time. Two-sample, unpaired t-tests were therefore conducted in which the average frequency of each concept's appearance in European IS research articles during a particular year were compared against their analogous frequencies of appearance in North American IS research articles during the same year. Due to the large number of hypothesis tests, the observed p-values were corrected using the Bonferroni method according to the number of statistical tests conducted for each year of the analysis [21, 22]. The results obtained from the overall analysis are illustrated in Figure 4 below, which depicts third-order polynomial fit lines for the concept frequency t-tests at Bonferroni-corrected significance levels of 0.05, 0.01, and 0.001. The R^2 values for these fit lines were very large at 0.87, 0.83, and 0.82, respectively.

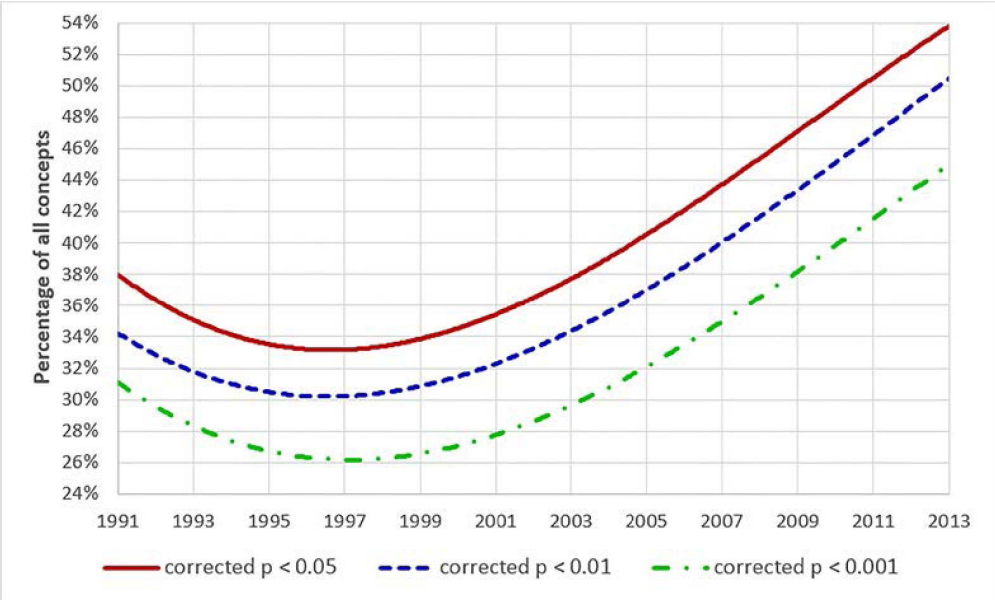

Figure 4. Percentage of concepts with statistically different frequencies of appearance in European vs. North American IS research from 1991 to 2013.

As shown in the figure above, the overall percentage of concepts appearing in European IS research whose frequency of usage differed significantly from North American IS research declined steadily from the early-to-mid 1990s, indicating that the topics being addressed in IS research on the two continents were converging. Beginning in the mid-to-late 1990s, however, this pattern experienced a sharp reversal, and the data indicate that European IS research and North American IS research have been addressing an increasingly divergent set of topics since that time. The consideration of all of the concepts which appeared in the two corpora may be inferentially misleading, however, since many concepts may appear very infrequently or sporadically over time. If such concepts happen to appear exclusively or primarily in either European or North American journals, then the outcomes of their associated t-tests could be expected to introduce a great deal of noise into the analytical findings. For this reason, the analysis described above was repeated using a Bonferroni-corrected significance level of 0.001, with the data first being subdivided into quartiles according to the frequency with which each concept appeared in its respective corpus during each year. The results of this analysis are illustrated in Figure 5 on the following page.

The figure depicts third-order polynomial fit lines for the concept frequency t-tests, after the data were subdivided into quartiles according to the concepts' yearly frequencies of appearance. Although the R^2 values for the fourth, third, and second quartile fit lines were very large and highly reliable at 0.82, 0.84, and 0.88, respectively, the R^2 value for the first quartile fit line was only 0.05, thus revealing that rare concepts (i.e., concepts which appear infrequently or sporadically in the corpus) did indeed add a great deal of contamination to the original analysis. By contrast, the concept frequency data in the upper three quartiles follow the same general pattern reported in the original analysis, albeit in a less alarming manner. Specifically, the percentage of non-rare concepts whose frequency of usage differed significantly between European and North American IS research declined steadily from the early 1990s until approximately 1997, at which point the topics being discussed in IS research on the two continents were the most similar (i.e., only 3-4% of all of the non-rare concepts used in IS research in 1997 differed in their frequency of appearance in leading European and North American IS research journals). Beginning in 1998, however, European and North American IS research began to diverge, with research published on the two continents thereafter addressing increasingly distinct sets of topics. Put another way, since 1997 European IS research and North American IS research have become more and more distinct from one another, and the data suggest that this intercontinental distinctiveness will continue to grow for many years to come. Just as tectonic geological forces are causing the continents themselves to slowly drift apart, so too, apparently, is there a growing separation in the nature and character of European and North American IS research.

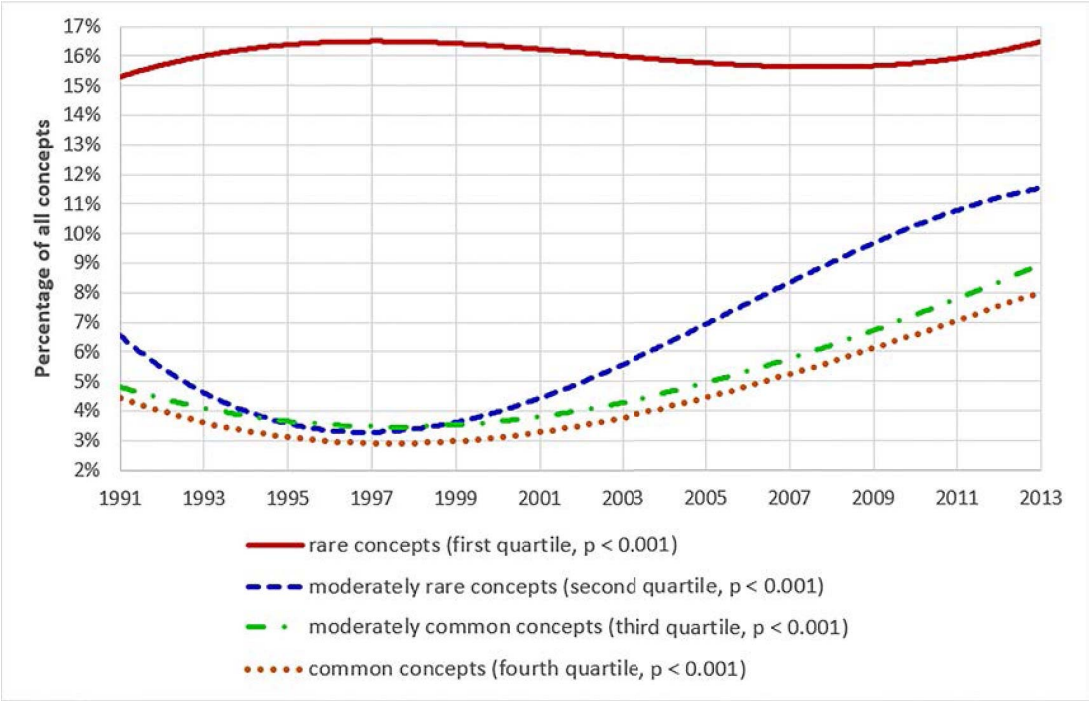

Figure 5. Percentage of concepts with statistically different frequencies of appearance in European vs. North American IS research from 1991 to 2013, arranged by frequency quartile.

5. Conclusion

The findings show that the technique we introduce, *concept frequency analysis*, is a viable research tool that can help researchers to address grand research questions which involve the analysis of large corpora of text. When applied to the possible divide between, or convergence of, North American and European IS research, this technique revealed significant structural differences (conceptual density, words per article, concepts per article) between IS research affiliated with each of these continents. While the conceptual density of research affiliated with each one of these continents seems to be slowly converging, differences in the concepts discussed in these continents seem to have been growing since the mid-1990s. Hence, the answer to the question "Is the world of IS research flat?" is mostly no. On the one hand, this may be considered undesirable, since it defies global trends of convergence [3, 4]. On the other hand, it may be beneficial since it provides additional space for the discussion of different concepts, and makes IS a broader and richer field of inquiry [23, 24].

The interpretation of the presented results should take into account two limitations. First, we only used approximately 10 million concept labels. While we believe that this set included all major concepts, it is likely that additional, less common concepts were not considered. Future research may develop ways of extending the concept pool used herein. In addition, we only considered four leading European and North American IS journals. Future research can extend the scope of our study, and include a broader set of journals representing these continents.

5.1 Concluding remarks

The rapidly growing divergence between European and North American IS research is very interesting when considered in light of theories of globalization and the development of a single global culture. Whereas such theories would naturally predict convergence in the topics being examined in European and North American IS research over time, the data suggest that the opposite is true. Put differently, although research, popular media, and other accounts have documented many ways in which the cultures of Europe and North America are becoming increasingly similar, this pattern does not appear to hold in the context of the information systems field. Instead, IS research on each of these two continents is, in actuality, becoming more and more culturally distinct.

6. References

[1] Isaacson, W., Einstein: His Life and Universe, Simon & Schuster, New York, NY, 2008.

[2] Chong, A., and Gradstein, M., "Is the World Flat? Country- and Firm-Level Determinants of Law

Compliance", Journal of Law Economics & Organization, 27(2), 2011, pp. 272-300.
[3] Feiock, R.C., Moon, M.J., and Park, H.J., "Is the world "flat" or "spiky"? Rethinking the governance implications of globalization for economic development", Public Administration Review, 68(1), 2008, pp. 24-35.
[4] Mithas, S., and Whitaker, J., "Is the world flat or spiky? Information intensity, skills, and global service disaggregation", Information Systems Research, 18(3), 2007, pp. 237-259.
[5] Carey, S., The Origin of Concepts, Oxford University Press, Oxford, UK, 2011.
[6] Murphy, G., The Big Book of Concepts, MIT Press, Cambridge, MA, 2004.
[7] Lawrence, S., and Margolis, E., Concepts: Core Readings, MIT Press, Cambridge, MA, 1999.
[8] Crystal, D., Dictionary of Linguistics and Phonetics (6th ed.), Wiley-Blackwell, Hoboken, NJ, 2008.
[9] Cruse, A., Meaning in Language: An Introduction to Semantics and Pragmatics, Oxford University Press, Oxford, UK, 2011.
[10] Manning, C.D., and Schütze, H., Foundations of statistical natural language processing, MIT Press, Cambridge, MA, 1999.
[11] Bohannon, J., "Google Books, Wikipedia, and the Future of Culturomics", Science, 331(6014), 2011, pp. 135.
[12] Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, T.G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Aiden, E.L., "Quantitative Analysis of Culture Using Millions of Digitized Books", Science, 331(6014), 2011, pp. 176.
[13] Soper, D.S., and Turel, O., "An n-Gram Analysis of Communications: 2000-2010", Communications of the ACM, 55(5), 2012, pp. 81-87.
[14] Soper, D.S., and Turel, O., "Who Are We? Mining Institutional Identities Using n-grams", 45th Hawaii International Conference on System Sciences (HICSS), 2012
[15] Soper, D.S., Turel, O., and Geri, N., "The Intellectual Core of the IS Field: A Systematic Exploration of Theories in Our Top Journals", 47th Hawaii International Conference on System Sciences (HICSS), 2014
[16] Ferratt, T.W., Gorman, M.F., Kanet, J.J., and Salisbury, W.D., "IS Journal Quality Assessment Using the Author Affiliation Index", Communications of the Association for Information Systems, 19, 2007, pp. 710-724.
[17] Rainer, K., and Miller, M., "Examining differences across journal rankings", Communications of the ACM, 48(2), 2005, pp. 91-94.
[18] Wikimedia Foundation, "Wikipedia, The Free Encyclopedia", in (Editor, 'ed.'^'eds.'): Book Wikipedia, The Free Encyclopedia, Wikimedia Foundation, Inc., San Francisco, CA, 2014
[19] Keller, J., "Is Wikipedia a World Cultural Repository?": The Atlantic, Atlantic Media Company, Washington, DC, 2011
[20] Words Worldwide, Word List of US-UK Spelling Variants, Words Worldwide Limited, Newcastle upon Tyne, UK, 2009.
[21] Dunn, O.J., "Estimation of the medians for dependent variables", The Annals of Mathematical Statistics, 30(1), 1959, pp. 192-197.
[22] Dunn, O.J., "Multiple comparisons among means", Journal of the American Statistical Association, 56(293), 1961, pp. 52-64.
[23] Baskerville, R., and Wood-Harper, A.T., "Diversity in information systems action research methods", European Journal of Information Systems, 7(2), 1998, pp. 90-107.
[24] Robey, D., "Research commentary: Diversity in information systems research: Threat, promise, and responsibility", Information Systems Research, 7(4), 1996, pp. 400-408.