2012 45th Hawaii International Conference on System Sciences

Who Are We? Mining Institutional Identities Using *n*-grams

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

Abstract

Disciplines and organizations alike can be defined by the text they produce, the topics they discuss, and the language they employ. Analyzing such large amounts of text is challenging, but is nevertheless needed because it can help stakeholders to understand key themes in, and the evolution of their corporate or disciplinary identity. N-gram analysis is a leading text-mining technique that can be leveraged for this purpose. In this manuscript we present the development and demonstrate the potential utility of an n-gram analysis tool. We focus on revealing several aspects of the identity of an academic journal, namely Communications of the ACM (CACM), through the analysis of over 14 million unique n-grams and their relative frequencies. The results of the study imply that n-gram analyses may be a key tool in resolving the IS identity crisis. Implications for research and practice are discussed.

1. Introduction

Institutions of all kinds, be they companies, governments, academic institutions, or other organizations, have distinguishing characteristics that make them unique in some way. For many of these institutions, a deep, tacit understanding of the institution's identity is difficult to acquire, especially in cases when the institution has a long history, large geographic footprint, or regular turnover among its employees or members. Decision-makers within these institutions will inevitably struggle to develop comprehensive knowledge of the cultural themes and historical concepts that have shaped their institutions over time. These themes and concepts may be embedded within the archived artifacts and documents produced by an institution throughout its history, but the sheer volume of such materials may not be digestible by decision-makers who are burdened with myriad other responsibilities. Without a solid and detailed understanding of her institution's identity, how can we reasonably expect a decision-maker to make fully-informed choices on behalf of her institution?

Many academic disciplines also suffer from this phenomenon, with the problem being especially pronounced in the information systems (IS) discipline [2]. A never-ending influx of new technologies, buzzwords, and trends has produced an environment characterized by fashion waves and constant change, and the resulting instability has made it difficult for the IS discipline to establish a stable identity for itself in the corporate and academic worlds [1, 3, 18, 33, 35, 36]. As archived institutional artifacts, the articles published in IS journals and other outlets reflect and codify the evolving identity of the discipline. Unfortunately IS academics and professionals cannot reasonably hope to digest all of the literature that has ever been published in the history of their discipline. By leveraging a computational method known as *n*-gram analysis, however, it may be possible to unlock some of the secrets embedded within this vast collection of institutional artifacts, and thereby gain insights into the culture, history, and evolution of the discipline that might otherwise remain shrouded in mystery. Put another way, if the history and identity of the IS discipline are encoded within the pages of its journals, then systematically analyzing the text published in those journals may yield a better understanding of where the discipline has been, where it is today, and where it might be heading in the future. Such analyses may, in fact, prove instrumental in understanding IS fashion waves and resolving the IS identity crisis [1]. After all, "We are what we write, we are what we read, and we are what we make of what we read" [4].

In the pages that follow, we present a culturomic[1] analysis [6] of a leading computing journal – *Communications of the ACM* (CACM). This widely-read journal provides an equitable mix of practitioner and research-oriented perspectives, and hence nicely reflects the different facets of the IS discipline as a whole. In conducting this analysis, we provide an example of how natural language processing [20] can be used to quantitatively explore the identity and

[1] "Culturomics" is an emerging field in which high-volume data collection and analytics are used to computationally study human culture.

978-0-7695-4525-7/12 $26.00 © 2012 IEEE
DOI 10.1109/HICSS.2012.642

culture of the discipline over time through a computational examination of its document artifacts. This work was largely inspired by Google's "*n*-grams" project, which was described in detail in a recent issue of the journal *Science* [25]. In natural language processing, an *n-gram* can be thought of as a sequence of words of length *n* which is extracted from a larger sequence of words [23]. Google's *n*-grams project was undertaken to support the quantitative study of cultural trends using combinations of words (*n*-grams) that appear in a corpus of millions of books. The central theory underlying Google's project is that when taken together, the words appearing in a large corpus of text speak to the identity and culture of a society at the point in time when that text was written [7]. By computationally analyzing those words, it thus becomes possible to study cultural evolution over time.

In this paper, we contend that a similar analytical approach and theoretical orientation can be usefully applied to other corpora – such as organizational artifacts and the publications of academic disciplines – in order to gain a better understanding of the key trends, cultural indicators, and evolutionary changes that have occurred in those institutions over time. With a view toward portraying the power and flexibility of this approach, we present several informative findings that emerged from our *n*-gram analysis of CACM. While these are by no means the only findings that might be extracted from our vast CACM *n*-grams database, they nevertheless demonstrate the viability of the approach for institutional identity data mining. Our efforts here are focused on only a single IS journal, but we hope that our research will engender future studies that evaluate and compare the cultures and identities of a large basket of IS journals, thereby providing a deeper understanding of the history and evolution of the discipline over time. Such analyses may contribute greatly to our knowledge about IS fashion waves, and to efforts aimed at resolving the IS identity crisis.

2. Method

In this study, we have taken an exploratory rather than a confirmatory approach to examine how the identity of the IS discipline has evolved in recent years. The observations and measurements reported in this paper represent the first step in an inductive theory building process oriented toward understanding the identity and culture of the field [21, 27, 32]. Since data form the foundation of the case-based inductive approach [16], we began by constructing a corpus containing the complete text of every article published in CACM between 2000 and 2010[2]. CACM was chosen for our analysis because of (1) its status as a leading computing journal [30, 22], (2) its dual focus on research and practice, and (3) because its articles are made available in a machine-readable format. In total, our corpus contained 3,367 articles which together comprised more than 8.1 million words. To put the size of the corpus in perspective, consider that if you were to spend 40 hours per week reading CACM, you would need more than four months to read every article that was published by the journal between 2000 and 2010.

With our corpus complete, we next constructed a custom software system to tokenize the text of each article into a series of *n*-grams. René Descartes' famous phrase *cogito ergo sum* [14], for example, can be subdivided into three *1*-grams (*cogito*, *ergo*, and *sum*), two *2*-grams (*cogito ergo*, and *ergo sum*), and one *3*-gram (*cogito ergo sum*). As this example illustrates, the number of *n*-grams that can be extracted from a large corpus of text greatly exceeds the number of words in the corpus itself. This situation presents serious scaling and performance implications for a corpus containing millions of words, so our analysis was constrained to include *n*-grams with a maximum length of $n = 4$.

To address the challenges presented by punctuation, we adopted a technique which was successfully used by the developers of Google's *n*-gram project for digitized books [25]. Using this approach, most punctuation marks are treated as separate words during the *n*-gram tokenization process. The phrase "Wherefore art thou Romeo?" would, for example, be tokenized as five words:

Wherefore	art	thou	Romeo	?

There are a few notable exceptions to this rule, such as currency symbols, decimal components of numbers, and apostrophes indicating possessive case. A term such as "$5.95", for example, would be treated as a *1*-gram, while "Avogadro's number" would be treated as a *2*-gram.

In designing our custom *n*-gram analysis software, we also made the decision to ignore case in the construction of the CACM *n*-grams database. Had we retained case sensitivity, then a term such as "Cloud Computing" would be treated as distinct from the term "cloud computing". While ignoring case vastly reduced the potential number of *n*-grams that the system might encounter, it also had a few negative implications for search specificity. Without case sensitivity, for example, the term "IT" (in reference to *Information*

[2] Only the actual text of each article was included in the database – trailing matter such as acknowledgements and references were excluded.

Technology) would be considered identical to the word "it". Despite this drawback, we concluded that the overall benefits to our project of ignoring case outweighed the costs.

Broadly speaking, our *n*-gram analysis of CACM is predicated on the idea that the degree of currency or interest in a particular concept is reflected in the relative frequency with which that concept appears in the journal over time. For example, if the *n*-gram "e-commerce" was mentioned 273 times in 2000, but only 23 times in 2010[3], then we might infer that the concept of "e-commerce" has become less popular or perhaps less influential within the journal over time. It was therefore necessary to compute the frequency with which every *n*-gram in the corpus appeared in CACM during each year of the analysis. The direct comparison of *n*-gram frequencies would be misleading, however, because doing so would ignore potential differences in the number of words published by the journal from year to year. It was therefore necessary to calculate relative frequencies for each *n*-gram by dividing their respective raw frequency counts by the total number of words appearing in the corpus during a given year. This approach yielded a standardized measure of frequency which would allow valid comparisons to be made between *n*-grams over time [25]. The standardized frequency values resulting from this process thus indicated how often a particular *n*-gram appeared in CACM during a particular year, relative to the total quantity of text published in the journal during that year.

The result of all of these data extraction and processing tasks was a vast database containing more than 14.5 million unique *n*-grams. Since a standardized frequency measure for each of these *n*-grams was computed for each of the 11 years of the analysis, the final dataset contained more than 160 million rows of data. From this collection we then selected the 1 million unique *n*-grams which exhibited the most absolute change over time, reasoning that the frequencies of less interesting *n*-grams such as "how" and "the" would remain relatively stable from year to year. Finally, we constructed a custom web-based system that enabled us to query, graph, and explore our CACM historical *n*-grams database. This system allowed us not only to plot and analyze multiple *n*-grams simultaneously, but also to combine related search terms into a single result. For example, the search phrase "cellphone + cellphones, smartphone + smartphones" would produce a graph containing two lines, one representing the combined frequencies of the terms "cellphone" and "cellphones" over time, and one representing the combined frequencies of the terms "smartphone" and "smartphones" over time. Interested readers may use our CACM *n*-grams tool to conduct their own analyses by navigating to http://www.invivo.co/ngrams/cacm.aspx.

Before presenting our findings, a brief discussion of semantic error may be useful. Most words and phrases in natural human languages can have multiple meanings depending upon the context in which they appear. This fact has the unfortunate side effect of introducing a degree of semantic error into all methods of culturomic analysis, including the *n*-gram approach used herein. For example, a search for the *n*-gram "PC" in reference to *Personal Computer* would be indistinguishable from a search for the *n*-gram "PC" in reference to *Politically Correct*. The relative frequency value for every *n*-gram thus contains an element of residual error that reflects the distance between the observed frequency value and the semantically "true" value. Virtually all forms of quantitative data contain measurement error, and *n*-gram frequencies are no exception. Since the degree of semantic error from frequency to frequency is random rather than systematic, however, the expected value of the residual error for any randomly selected *n*-gram frequency is zero [19]. This implies that the sum of squares of *k* semantic error residuals divided by their associated variance will be approximately chi-square distributed with *k*-1 degrees of freedom. This information lays the foundation for bridging the divide between *n*-gram analyses and traditional methods of statistical analysis and prediction.

Further, it important to note that several factors can produce changes in the relative frequency with which a particular *n*-gram appears in the corpus over time. Among these are: (1) factors specific to the CACM journal, (2) factors specific to the IS discipline, and (3) linguistic factors specific to the *n*-gram under consideration. Efforts aimed at uncovering the antecedents of an observed change in *n*-gram frequencies may therefore benefit from parceling the frequency changes according to these factors.

3. Findings

While we could not reasonably hope to identify and describe *all* of the ways in which the IS discipline has evolved in a single conference paper that relies on a single journal as a data source, we nevertheless sought to explore and document some of these vicissitudes in order to provide a point of embarkation for future research, and to demonstrate the value of *n*-gram culturomic analyses. We begin by presenting a few findings related to the evolution of CACM itself.

[3] These are the actual *n*-gram frequencies for "e-commerce" during 2000 and 2010.

3.1 The evolution of a journal

Our *n*-gram analyses indicate that the focus of CACM has changed markedly in recent years, and future research should assess the extent to which these changes are generalizable across other journals in the computing field. Specifically, the content of the journal is becoming more technical and scientific, as indicated by strong growth between 2000 and 2010 in the relative frequencies of *n*-grams such as "computer science" (+616%), "source code" (+400%), "algorithm" (+415%), "theory" (+204%), and "research" (+166%). Conversely, the focus of the journal appears to be shifting away from business issues and organizational systems. With respect to the former, interest in business-related issues has been declining steadily since 2000, as evidenced by substantial decreases in the appearance of *n*-grams such as "business" (-226%), "user" (-241%), "customer" (-176%), and "market" (-205%). Interest in organizational systems within CACM peaked around 2003, but has also declined sharply in the intervening years. Evidence of this can be seen in the waning relative frequencies of *n*-grams such as "crm+customer relationship management" (-747%), "decision support system" (-644%), and "erp+enterprise resource planning" (-460%).

From a broader perspective, our metadata revealed several striking, large-scale structural changes in the journal between 2000 and 2010. During this time, CACM published an average of about 745,000 words per year divided among an average of 306 articles per year. These averages, however, obscure the underlying trends, which indicate that both the number of articles published per year and the total number of words published per year have been growing markedly. These trends, which are depicted in the figure below, show that the journal has been responsive to calls within the discipline to increase the number of articles published [13].

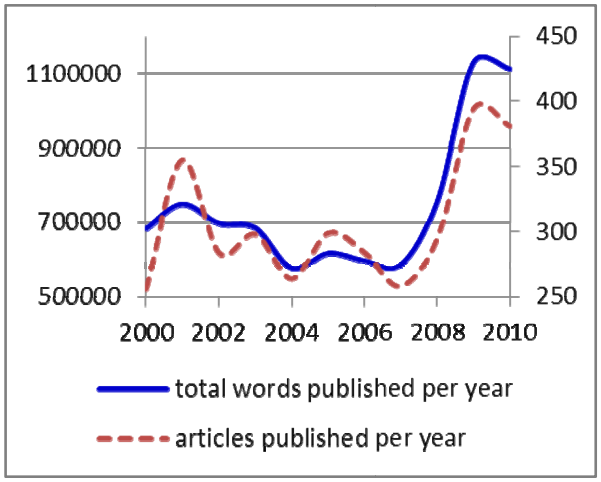

Figure 1. CACM publication volumetrics.

At present, we have no evidence to suggest that these findings extend beyond CACM to represent a more general trend among journals in the discipline – analyses of other top IS journals are clearly needed in order to provide a more complete picture.

3.2 Changes in writing style

The style of the writing in CACM evolved substantially between 2000 and 2010. Authors of CACM articles seem to be transitioning away from the traditional scientific style of writing, and instead are adopting a less formal, more personal tone. Evidence of this change in style can be seen in the increasing use of words that refer directly to the author(s) of an article, such as "I" (+143%) and "we" (+137%), and in the increasing frequency with which authors speak directly to their readers by using words such as "you" (+222%) and "your" (+205%). We, too, have found ourselves adopting this more personal style when writing the current paper, and hope that future research will reveal whether this trend toward informality extends to other leading IS journals. It would also be interesting to determine whether this is an IS phenomenon, or a more general research phenomenon (*i.e.*, on average, it may be that all academic journals are becoming less formal).

The usage of gender-related terms has also been changing in CACM. In the aggregate, masculine pronouns such as "he", "his", and "him" appear in the journal 277% more often than feminine pronouns such as "she", "hers", and "her". What's more, this gap has widened from 190% in 2000 to more than 290% in 2010, as shown in Figure 2 below.

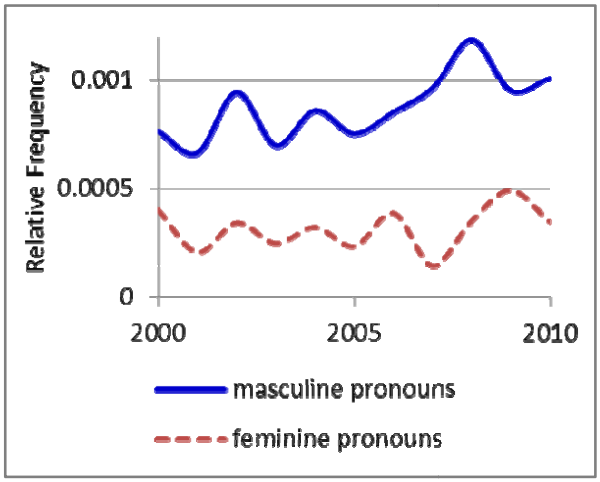

Figure 2. The gender gap in CACM.

One possible explanation for this phenomenon is that it reflects the gender gap between male and female computing professionals, which is wider today than it has been at any time in the past 25 years [26]. Because IS has traditionally been male dominated, masculine values often predominate in the discipline [29, 31]. Our analysis provides further empirical support for this argument, since it shows that the gender gap is also entrenched in the way we write within the profession.

3.3 The systems development life cycle

The systems development life cycle (SDLC) has been one of the most enduring components of the information systems discipline [24]. For the four principal phases of the SDLC (planning, analysis, design, and implementation), current industry standards suggest that approximately 15% of the attention and resources available for a systems development project be allocated to planning, 20% be used for analysis, 35% for design, and 30% for implementation [12]. To what extent, then, does interest in these four SDLC phases within the pages of IS journals mirror the level of interest recommended by industry standards? We cannot answer this question without considering all IS journals, but we can examine it through the lens of CACM. To answer this question within this scope, an *n*-gram analysis was used to compute the average frequency with which each phase of the SDLC was mentioned in CACM between 2000 and 2010. Dividing the value for each phase by the overall sum of the average frequencies yielded the relative frequency distribution for CACM shown in Table 1 below.

Table 1. Interest in the phases of the SDLC: CACM vs. industry standards

SDLC Phase	Level of Interest	
	Industry Standard	CACM
Planning	15%	8%
Analysis	20%	29%
Design	35%	46%
Implementation	30%	17%

If we accept current industry standards as canonical, then the values in the table suggest that CACM underemphasizes the planning and implementation phases of the SDLC, while overemphasizing analysis and design. Both sources, however, would seem to agree in principle that design deserves the most attention during a systems development project, while planning deserves the least attention. The overall discrepancies between these two sources also raise another interesting possibility – if in the aggregate IS journals accurately reflect the interests and behaviors of information systems professionals, then it may be prudent to investigate the apparent misalignment between recommended industry standards and real world practice.

3.4 The impact of special sections and issues

Most IS journals occasionally publish a special section or special issue that focuses on a specific topic of interest to the editorial board. But do these special sections and issues engender long-term interest in the topic being addressed, or is their effect more fleeting? To gain insight into these questions, we selected three topics that have been the focus of special sections in past issues of CACM: democracy [8], DRM / digital rights management [10], and spyware [9]. Our only criterion in selecting these particular special section topics was that they were published during the early years of our analytic period, thus allowing any long-term effects to be identified. The results of our *n*-gram analysis of these three special section topics are shown in Figure 3 below:

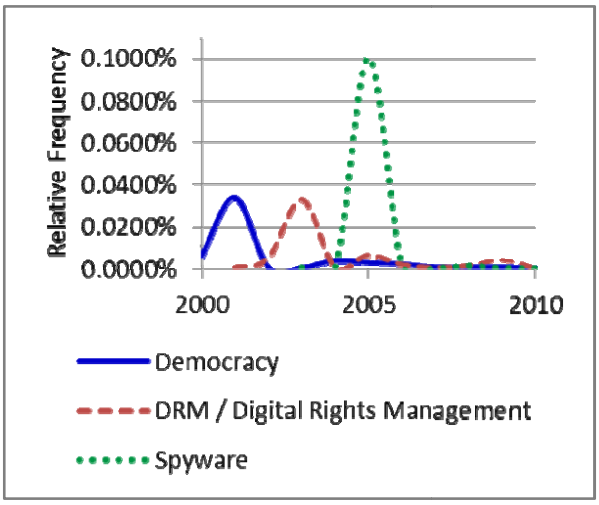

Figure 3. The effect of special sections on long-term topic interest

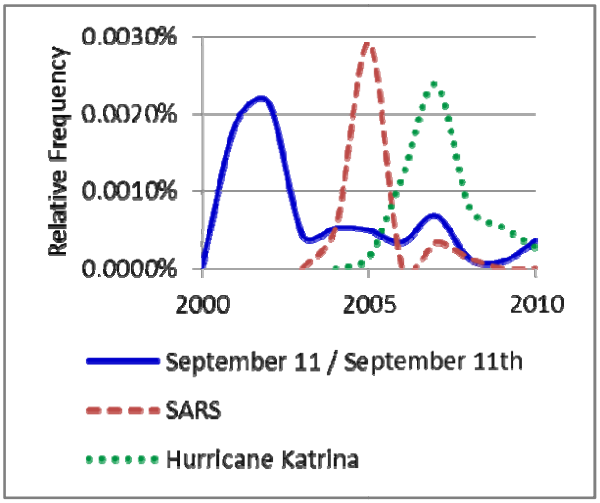

Figure 4. Responsiveness of CACM to major world events

As shown in the figure, a special section on a given topic produces a spike in interest in that topic during the year in which the special section is published. This elevated interest is sustained for a very short time before rapidly declining, eventually returning to a near-zero steady state. Thus, although they can be expected to increase the visibility of a topic in the short-term, special sections do not seem to engender any sustained long-term interest in the topics that they address, at least within CACM. Additional research will be needed to determine whether this observation holds true for other IS journals, and whether special topic issues have influence beyond their respective journals (*i.e.*, they are cited or discussed in other journals).

3.5 IS journals and major world events

Do major world events influence the information systems discipline? To answer this question, we conducted an *n*-gram analysis within the scope of CACM that included three different types of world events – a natural disaster (Hurricane Katrina), a terrorist attack (September 11th), and a health crisis (the 2003 SARS outbreak). Figure 4 below summarizes the results of this analysis:

As shown in the figure, major world events exert a strong influence on CACM, and a common pattern exists with respect to the nature of this influence. Specifically, a major world event will first appear in the pages of CACM shortly after the event occurs, and discussion of that event grows rapidly for a short time thereafter. This finding indicates that CACM is not insulated from major world events, but rather that it embraces such events and actively contributes to their discussion in the global forum. After a period of 1-2 years, however, the journal's interest in a major event declines sharply. Nevertheless, even after suffering this precipitous drop in interest, major world events tend to be mentioned occasionally in the journal for several years to come. Additional research will be needed to establish whether this pattern holds for other leading IS journals and for the discipline in general.

3.6 Technology preferences

If the articles published in IS journals truly reflect the state of the art in the computing profession, then an *n*-gram analysis focused on specific technologies should prove useful in evaluating differences between the technological interests and preferences of IS professionals and those of the computing public at large. We thus compared the CACM *n*-gram frequencies of different web browsers and operating systems in 2010 with the market shares of those products among the general public during the same year [28]. The results of this analysis are shown in Figures 5 and 6 below:

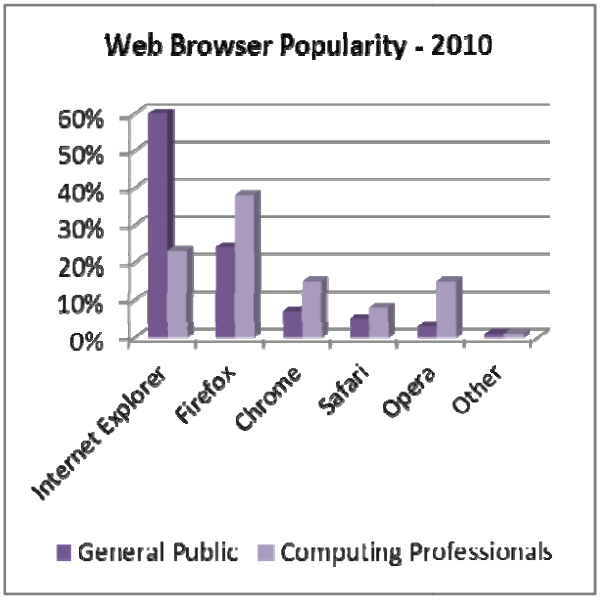

Figure 5. Comparative popularity of web browsers

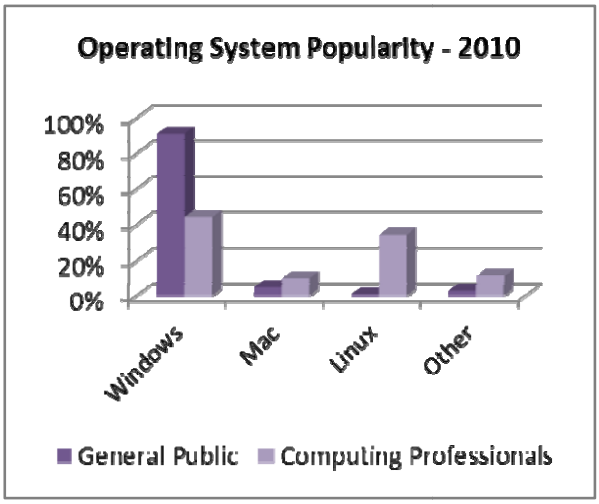

Figure 6. Comparative popularity of operating systems

The figures above indicate that the preferences and interests of computing professionals differ markedly with respect to web browsers and operating systems from those of the public at large. More specifically, there appears to be a greater degree of diversity among the preferences of computing professionals regarding these technologies, while the usage patterns among the general public are much more homogeneous. One plausible explanation for this phenomenon is that computing professionals may have a more nuanced understanding of the advantages and disadvantages of the various technology options available to them, and are hence more likely to orient their preferences toward the technologies that best service their distinctive needs.

3.7 Technology life cycle

Does an IS journal's level of interest in a particular technology proxy the location of that technology along its life cycle? To investigate this issue, we conducted an *n*-gram analysis of three Apple technologies which are arguably at different points in their respective life cycles: the iPod, the iPhone, and the iPad. Figure 7 below summarizes the results:

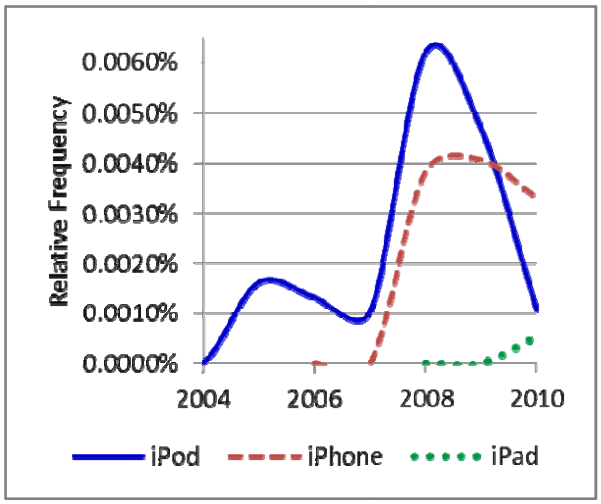

Figure 7. Using *n*-grams to assess the rise and fall of different Apple technologies

As shown in the figure, the frequency with which a particular technology appears in the pages of CACM appears to echo the location of that technology along its own unique life cycle. Interest in the iPod within CACM, for example, grew sharply between 2004 and 2005, corresponding to a 500% increase in sales during that time period [15]. After the rise of the iPhone between 2007 and 2008, however, interest in the iPod began to wane, as the ongoing incorporation of digital music capabilities into smartphones (such as the iPhone) made the iPod an increasingly redundant technology. More recently, interest in the iPhone has also begun to decline, while at the same time interest in Apple's latest device – the iPad – is beginning to rise. Taken together, we believe that these results demonstrate the viability of using *n*-gram analyses for the purpose of studying technological evolution over time. In managerial contexts, similar analyses can be used to study the evolution of organizational issues over time, *e.g.*, the focus of customer complaints, or issues employees discuss in company blogs.

3.8 Comparing the prominence and impact of different theories

Finally, we wish to illustrate how *n*-grams might be used to quantitatively identify which theories have been most important to the IS discipline, at least from a CACM readership perspective. While we acknowledge that the scope of this question is far too grand to address using data from a single IS journal, we nevertheless want to demonstrate the potential of *n*-grams to contribute within this domain of inquiry. We therefore conducted an *n*-gram analysis that compared two prominent behavioral theories – the Technology Acceptance Model [11] and the Theory of Reasoned Action [17] – to serve as an exemplar of the usefulness of the *n*-grams approach for answering such questions. The results of this analysis are shown in Figure 8 below:

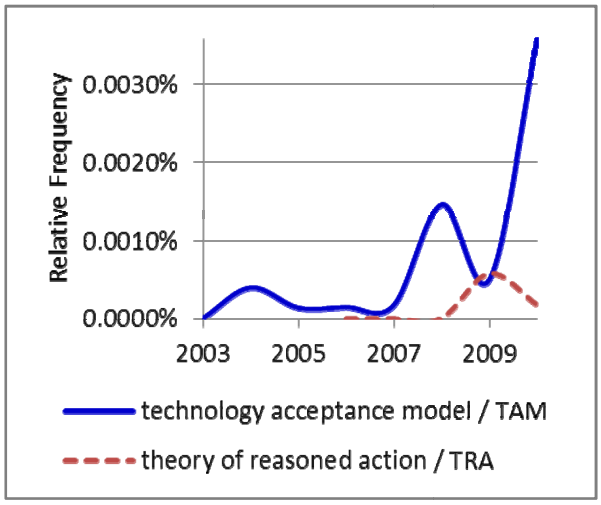

Figure 8. Using *n*-grams to compare the prominence and impact of IS theories

As shown in the figure, applying the *n*-grams technique to these two theories reveals that the technology acceptance model has played a much more prominent role in CACM since the turn of the century than has the theory of reasoned action. From a comparative perspective, the relative total impact of a given theory can be quantified simply by computing the area under its respective curve. In the figure above, for example, the total area under the "technology acceptance model" curve is approximately 450% greater than the total area under the "theory of reasoned action" curve, thus providing a quantitative measure of the degree of difference between each theory's impact within CACM. What's more, using the *n*-grams approach also reveals *when* each theory was exerting its greatest influence on the discipline. Applying *n*-gram analyses to a broad collection of IS theories and journals would, we believe, make for a fascinating study.

4. Summary, Limitations, and Implications

In this paper we demonstrated how *n*-gram analyses can be used to study different facets of the history, culture, and evolution of academic journals over time with a view toward gaining insights into the identity of the underlying discipline. Using CACM as an exemplar, we presented an array of analyses that together demonstrate the vast potential of the *n*-grams method for culturomic inquiry into questions of institutional identity. When combined with the results of similar future studies, the observations and measurements reported here can meaningfully contribute to a larger theory building process focused on unlocking the identity and culture of the IS discipline. This paper also stands as an example of how natural language processing can be used to quantitatively explore the identity and culture of virtually any text-producing institution by computationally examining its document artifacts, and opens the door for interorganizational and interdisciplinary analyses based on the *n*-grams method.

4.1 Limitations

Although we presented a diverse assortment of analyses, we acknowledge that the results reported here represent but a tiny fraction of the analyses that potentially could be conducted using the *n*-grams approach. We therefore invite interested readers to conduct their own analyses of CACM using our *n*-grams tool, which is available online at: http://www.invivo.co/ngrams/cacm.aspx. Researchers must be cautious when interpreting these *n*-gram analyses, however, since there are many factors which might contribute to changes in *n*-gram frequencies over time.

The *n*-gram method itself is also limited inasmuch as it does not fully capture the semantic context of the various concepts and themes that researchers may be interested in investigating. Fortunately, the computational tools and methods used for natural language processing are evolving rapidly, and these tools and methods will undoubtedly prove instrumental in conducting semantic culturomic analyses after they have been given an opportunity to mature.

Since our results view the IS discipline through the narrow lens of a single journal – namely CACM – we encourage continued research that applies this approach to other IS journals. With respect to the IS discipline, we particularly encourage future research that compares and contrasts multiple leading IS

journals as well as journals in reference disciplines with a view toward quantitatively establishing the unique identity of each. Additionally, similar *n*-gram analysis tools can be built to consume and process data from organizational text repositories (e.g., an email server, a company's website, or customer forums).

4.2 Managerial implications

Recent findings reported in a longitudinal U.S. Department of Labor study indicated that college-educated workers will hold an average of 11 different jobs by the time they reach age 44 [5]. This alarming statistic has important ramifications for organizations, since it implies that skilled workers such as managers will, on average, hold their positions of leadership within an organization for only a few short years before moving on to greener pastures. With such a short tenure, managers will inexorably struggle to develop the sort of deep, tacit knowledge of their organizations that is so critical to effective strategic decision-making. If we believe that such knowledge can be found within the archived document artifacts that are produced by an institution over time, then *n*-gram analyses may prove to be an invaluable tool for sifting through and evaluating the content of these vast collections of archival documents, thus contributing to improvements in managerial decision-making.

Put another way, *n*-gram analyses in organizational contexts can produce actionable knowledge that might otherwise remain hidden. For example, if the phrase "no dental coverage" is observed to be increasingly discussed in employee forums, companies may choose to reevaluate their current dental coverage (or lack thereof). Alternatively, companies may evaluate corporate behavioral characteristics such as aggressiveness by comparing the frequencies with which defensive / passive terms and offensive / aggressive terms appear in internal emails. Future research should examine the utility of such analyses.

4.3 Research implications

We hope that this article serves as evidence for why leading IS journals should not be considered static and immutable, but rather should be seen as living entities that change, evolve, and respond to their environments over time. We therefore believe that analyses such as those described here should be conducted regularly in order to codify and record a journal's history, and help to establish and refine journal and academic disciplinary identities over time. When applied to a broad portfolio of journals, such periodic analyses would help readers, authors, and editors alike to better understand and compare journals more objectively. *N*-gram analyses may therefore prove to be a key tool in the ongoing effort to resolve the IS identity crisis [1], and perhaps even give us a brief glimpse of the directions in which the IS discipline may be moving in the future.

Speaking more broadly, we believe *n*-grams to be a powerful tool for gaining insights into textual data that might otherwise remain hidden. Although our analyses in this article were constrained to a single IS journal, the *n*-grams method itself is suitable for a wide array of research situations, and can be readily applied to virtually any large corpus of text-based data. Historical documents, for example, could be analyzed to identify long-hidden trends, fads, or modes of thought [7]. Software source code could be analyzed for style or efficiency. Blogs or tweets could be analyzed in order to produce a near real-time snapshot of the global consciousness. Clearly many possibilities exist, and when considered in this light, the potential for *n*-gram-based research seems almost limitless. From a theory-building perspective, such *n*-gram systems can be used as a basis for data mining procedures [34] that can reveal new associations between IS and organizational variables. They may also ultimately help us to better understand who we are, and to further decipher the identity of the IS discipline.

5. References

[1] R. Agarwal and H. C. Lucas, Jr., "The information systems identity crisis: Focusing on high-visibility and high-impact research", MIS Quarterly, 29 (2005), pp. 381-398.

[2] R. Agarwal and J. H. C. Lucas, "The Information Systems Identity Crisis: Focusing on High-Visibility and High-Impact Research", MIS Quarterly, 29 (2005), pp. 381-398.

[3] I. Benbasat and R. W. Zmud, "The identity crisis within the IS discipline: Defining and communicating the discipline's core properties", MIS Quarterly, 27 (2003), pp. 183-194.

[4] M. Bloomer, P. Hodkinson and S. Billett, "The Significance of Ontogeny and Habitus in Constructing Theories of Learning", Studies in Continuing Education, 26 (2004), pp. 19-43.

[5] BLS, *Number of Jobs Held, Labor Market Activity, and Earnings Growth Among the Youngest Baby Boomers: Results From a Longitudinal Survey*, Bureau of Labor Statistics, U.S. Department of Labor, Washington, D.C., 2010.

[6] J. Bohannon, "Google Books, Wikipedia, and the Future of Culturomics", Science, 331 (2011), pp. 135.

[7] D. Buscaldi, P. Rosso, J. M. Gomez-Soriano and E. Sanchis, "Answering questions with an n-gram based

passage retrieval engine", Journal of Intelligent Information Systems, 34 (2010), pp. 113-134.
[8] D. Crawford, "Editorial", Communications of the ACM, 44 (2001), pp. 5.
[9] D. Crawford, "Editorial pointers", Communications of the ACM, 48 (2005), pp. 5.
[10] D. Crawford, "Editorial pointers", Communications of the ACM, 46 (2003), pp. 5.
[11] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly, 13 (1989), pp. 319-340.
[12] A. Dennis, B. H. Wixom and R. M. Roth, *Systems Analysis and Design (4th ed.)*, Wiley, Hoboken, NJ, 2009.
[13] A. R. Dennis, J. S. Valacich, M. A. Fuller and C. Schneider, "Research Standards for Promotion and Tenure in Information Systems", MIS Quarterly, 30 (2006), pp. 1-12.
[14] R. Descartes, *Principia Philosophiae*, Kluwer Academic, Norwell, MA, 1644 [1984].
[15] S. Dowling, *Apple Reports First Quarter Results: Quarterly Revenue & Net Income Highest in Apple's History*, Apple Computer, Inc., Cupertino, CA, 2005.
[16] K. M. Eisenhardt and M. E. Graebner, "Theory Building From Cases: Opportunities and Challenges", The Academy of Management Journal, 50 (2007), pp. 25-32.
[17] M. Fishbein and I. Ajzen, *Belief, attitude, intention, and behavior: An introduction to theory and research*, Addison-Wesley, Reading, MA, 1975.
[18] G. Gill and A. Bhattacherjee, "WHOM ARE WE INFORMING? ISSUES AND RECOMMENDATIONS FOR MIS RESEARCH FROM AN INFORMING SCIENCES PERSPECTIVE", MIS Quarterly, 33 (2009), pp. 217-235.
[19] M. Grabe, *Measurement Uncertainties in Science and Technology*, Springer-Verlag, Berlin, 2010.
[20] G. Hripcsak, C. Friedman, P. O. Alderson, W. Dumouchel, S. B. Johnson and P. D. Clayton, "Unlocking clinical-data from narrative reports: A study of naturallanguage processing", Annals of Internal Medicine, 122 (1995), pp. 681-688.
[21] A. Lee, J. Liebenau and J. DeGross, "Rigor and relevance in MIS research - Introduction (Reprinted from Information Systems and Qualitative Research, 1997)", MIS Quarterly, 23 (1999), pp. 1-2.
[22] P. Lowry, D. Romans and A. Curtis, "Global journal prestige and supporting disciplines: A scientometric study of information systems journals", Journal of the Association for Information Systems, 5 (2004), pp. 29-75.
[23] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*, MIT Press, Cambridge, MA, 1999.
[24] M. M. Mantei and T. J. Teorey, "Incorporating behavioral techniques into the system development life-cycle", Mis Quarterly, 13 (1989), pp. 257-274.
[25] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak and E. L. Aiden, "Quantitative Analysis of Culture Using Millions of Digitized Books", Science, 331 (2011), pp. 176.
[26] T. J. Misa, *Gender Codes: Why Women Are Leaving Computing*, Wiley, Hoboken, NJ, 2010.
[27] M. D. Myers, "Qualitative research in information systems", MIS Quarterly, 21 (1997), pp. 241-242.
[28] NetMarketShare, *Usage Share Statistics for Internet Technologies*, Net Applications.com, Aliso Viejo, CA, 2010.
[29] F. Pratto, L. M. Stallworth, J. Sidanius and B. Siers, "The gender gap in occupational role attainment: A social dominance approach", Journal of Personality and Social Psychology, 72 (1997), pp. 37-53.
[30] K. Rainer and M. Miller, "Examining differences across journal rankings", Communications of the ACM, 48 (2005), pp. 91-94.
[31] M. F. Reid, M. W. Allen, D. J. Armstrong and C. K. Riemenschneider, "Perspectives on challenges facing women in IS: the cognitive gender gap", European Journal of Information Systems, 19 (2010), pp. 526-539.
[32] D. Remenyi and B. Williams, "The nature of research: Qualitative or quantitative, narrative or paradigmatic?" Information Systems Journal, 6 (1996), pp. 131-146.
[33] A. Serenko, M. Cocosila and O. Turel, "The State and Evolution of Information Systems Research in Canada: A Scientometric Analysis", Canadian Journal of Administrative Sciences-Revue Canadienne Des Sciences De L Administration, 25 (2008), pp. 279-294.
[34] G. Shmueli and O. Koppius, "Predictive Analytics in Information Systems Research", MIS Quartely, in print (forthcoming).
[35] M. J. Somers, "Using the theory of the professions to understand the IS identity crisis", European Journal of Information Systems, 19 (2010), pp. 382-388.
[36] H. Taylor, S. Dillon and M. Van Wingen, "Focus and diversity in information systems research: Meeting the dual demands of a healty applied science", MIS Quarterly, 34 (2010), pp. 647-667.