Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems

Kai R. Larsen⁎, David E. Monarchi, Dirk S. Hovorka, Christopher N. Bailey
Leeds School of Business, University of Colorado, Boulder, 419 UCB, Boulder, CO 80309, United States

A R T I C L E   I N F O

Article history:
Received 8 December 2006
Received in revised form 23 February 2008
Accepted 28 February 2008
Available online 6 March 2008

Keywords:
Latent Categorization Method
Unstructured data analysis
Organization of information systems
Research communities
Subfields
Research topics

A B S T R A C T

The Information Systems field is structured by the research topics emphasized by communities of journals. The Latent Categorization Method categorized and automatically named IS research topics in 14,510 abstracts from 65 Information Systems journals. These topics were clustered into seven intellectual communities based on publication patterns. The technique develops categories from the data itself, it is replicable, is relatively insensitive to the size of the text units, and it avoids many of the problems that frequently accompany human categorization. As such LCM provides a new approach to analyzing a wide array of textual data.

Contents lists available at ScienceDirect
Decision Support Systems
journal homepage: www.elsevier.com/locate/dss

1. Introduction

Since the inception of the field, Information Systems (IS) researchers have questioned the identity of the field and have attempted to map the field’s intellectual organization. More recently, researchers have proposed normative frameworks of what constitutes the “core” of IS research [7,50]. However, some researchers are concerned that such a normative perspective might be overly constraining and that the “the vibrancy of the information systems discipline lies in its porous boundaries, or the relatively large grey area that characterizes its domain, and the interests of the information systems research community” [52, p. 1]. This current research agrees with the view presented by DeSanctis [18] that the identity of IS lies in the action of systematic inquiry and the interactions within communities of practice. In this research, we are guided by Benbasat and Zmud’s [7] statement; “...the primary way in which a scholarly discipline signals its boundaries — and in doing so, its intellectual core — is through the topics that populate discipline-specific research activities” (p. 184). However in contrast to Benbasat and Zmud’s [7] normative approach, we argue that the identification of the field’s intellectual core is best achieved through empirical examination of what topics researchers actually investigate and publish in IS journals worldwide. We establish a historical view of the discipline [39] by determining the topics of intellectual interest to IS researchers over a 13 year period. We view IS as a “subject complex” which Apter [3] defines as “...an integration of parts of a wide range of other subjects in terms of a well developed body of ideas and techniques” (p. 99). Applied and theoretical sciences like IS, cybernetics, chemistry, and physics relate to a broad range of other subjects that can be studied in terms of them. As a discipline, IS considers both “what is” and “what may be” in its behavioral and design aspects [27,54], has multiple reference disciplines [32,34], and has multiple epistemological perspectives [45,49]. It is, therefore, unlikely that IS can revolve around a clearly defined and restricted set of topic areas or techniques because the subject matter itself is evolving and finding applications in new economic, personal, organizational, societal and cultural arenas.
The age and limitations of prior research combined with recent questions regarding the structure of the IS field suggest that an updated, accurate, and comprehensive overview would be of immense value to the field. In the absence of such a review, and due to the breadth and diversity of the IS field, researchers may not fully comprehend the scope of IS subfields, nor appreciate their colleagues’ research interests and communities. This lack of a comprehensive and replicable structure increases the difficulty of communicating within the field, especially in an international context. Recent research indicates that North American research can be differentiated from European research based on research topics, theories, and sources. Such differences support the argument that a comprehensive framework be based on representative journals from the whole field, rather than merely a narrow sample of North American publications. Vessey et al. [57, p. 170] captured this necessity in the statement:

We believe that the field should acknowledge not only those journals typically rated as being the top journals of our field, such as those examined in this study, but also journals that focus on areas such as system development, human–computer interaction, and so on. What is needed is a portfolio of journals that represents the field in its entirety and that can serve as guides not only to prospective authors but also to promotion and tenure committees.

The research presented here is a quantitative analysis of a larger dataset from a more extensive set of journals than has previously been examined. It is based on the perspective that scientific fields are organized and structured by the corpus of research problems and criteria for evaluation accepted by a scientific community of researchers. This research presents the structure of the IS field at two levels of analysis: the research topics that comprise the field, and the communities of interest that perform research and publish articles on specific sets of IS topics. In addition to those two levels, this research directly addresses the question of whether IS is composed of identifiable subfields as proposed by Banville and Landry. From our perspective, these subfields represent intellectual communities.

To identify the intellectual communities, the Latent Categorization Method (LCM) was used to cluster abstracts into automatically named topic areas. A total of 14,510 abstracts published in 65 journals was included, making it the most comprehensive sample of IS literature examined to date by an order of magnitude. This paper presents two new views of the IS discipline;

1) A quantitative and replicable analysis of the research topics published in the IS literature.
2) The intellectual communities of interest as represented by sets of interrelated topics, journal editors, reviewers, and contributors.

By using LCM to analyze research abstracts from a large set of IS journals, a comprehensive view of the scope and depth of research is developed. In addition, once a realistic view of the IS field gains acceptance, future research may be tailored to specific communities for the purposes of developing journal rankings, bodies of knowledge, and increasingly mature normative arguments.

This study introduces to Information Systems researchers the Latent Categorization Method (LCM), a new analytic technique that facilitates analysis of textual data that mathematically and computationally clusters and labels units of text, generally referred to as artifacts. LCM and related approaches have recently gained prominence in the IS field and are capable of handling large datasets. LCM aims to:

- develop a reproducible representation of artifacts (e.g., documents, interview data, survey data, etc.) and an approach to labeling that representation in a way that would (a) reduce... problems of human interpretation of the data; and (b) allow the application of quantitative techniques based on cardinal, rather than ordinal or nominal, data. Such an advance would offer an alternative as well as a complement to some existing methods for categorizing and labeling qualitative data.

Although human categorization of text data provides valuable insights into phenomena, this method is prone to two potential problems. First, the time required to perform human coding often limits the amount of data analyzed. Second, the analysis of text data is prone to variability of human categorization of research topics/keywords. As a computational technique, LCM is capable of analyzing very large datasets in a reproducible manner that is less susceptible to human interpretation bias.

Methods of semantic analysis, whether human or computationally based, are intended to determine the meaning of text-based data. Research articles use language to communicate perceptual experience and abstract concepts that map onto our inner or outer worlds. LCM and similar semantic analytic techniques view meaning as “almost entirely the relations that are represented and activated by words and collections of words.” From this perspective, to know what a word means is to possess the lexical network and “some set of techniques sufficient to attach the [word node] to the same experiences, objects or situations as are associated with it by other [speakers].” Thus, LCM does not provide a “better” technique for extracting meaning from text units; rather it provides a rigorous, reproducible and largely automatic method for comparing meaning in large text-based datasets.

Through a thorough review of the literature, we show that, although previous research in this area has been influential, LCM provides an opportunity to update and refine our conceptualization of the intellectual communities which form the IS discipline. Latent Categorization is a novel and groundbreaking method that provides reliability in that any researcher using this method with the same parameters will get identical results. Finally, we discuss the potential contributions of this new method to the field and discuss the implications of our analysis of intellectual communities within the IS field.

2. Structure of the IS field

2.1. Previous frameworks of the IS field

As noted by Vessey et al. [57, p. 130], “since its inception, the field has engaged in extensive self-examination.” This self-
examination has often been in response to criticism that the field is a fragmented adhocracy, and lacking a central theme. While one response to this criticism is a prescriptive view (e.g., [7]), another response is exemplified by the efforts to determine the structure of the field [e.g., 57]. Iivari et al. [30] noted that a considerable body of IS knowledge exists but that it is quite broad in scope and distributed over a large number of journals. They call for the IS research community to take stock of and organize its knowledge.

Previous studies (see Table 1) have provided immeasurable value to the IS field. Approaches have included historical documentation, the creation of keyword classification schemes, benchmarking, diversity assessments, frameworks, topic lists, taxonomies, social network overviews, and the examination of core vs. outlying topics.

Each of these studies examined a small selection of articles or abstracts from a subset of journals in the IS field. The limited samples and journal selection raise the question of whether they accurately represent the entire IS field or just mainstream research published in the highest ranked journals. For these reasons, the concept of the IS field as a fragmented adhocracy was introduced in a value-neutral fashion by Banville and Landry [4]. C. Banville and M. Landry, Can the field of MIS be disciplined? Communications of the ACM, 1989. 32(1): p. 48–60, but has later been cited in a negative context.

Studies to claim to map the intellectual organization of the IS discipline, one must assume that all subfields are well represented in just a few of the highest-ranking journals within a short time horizon. Swanson and Ramiller captured the common limitations to these studies when they commented on their own work that "while the ISR submissions stream gives one picture of the research activity in the field, work that some might label "IS" is almost certainly taking place out of the journal's sight" (p. 323). In addition, while North American journals dominate the rankings of IS journals, journals based outside of North America should be included to gain a complete view of IS as a discipline. It should also be noted that current ranking approaches will always discriminate against journals not part of the largest intellectual community, thereby invalidating arguments that only the top ranked journals in a field need be examined. Finally, none of these studies have examined DeSanctis' [18] communities of practice, an important oversight that this paper intends to address.

2.2. Subfields identification within the information systems discipline

Banville and Landry [4, p. 49] state that “it is generally agreed that our discipline is made up of what has (sic) been...
called subfields,” and Hirschheim and Klein [28] suggest that concise descriptions of the body of knowledge for a few subfields “could lift the discourse on the state of knowledge in the IS research community to a new level for both the internal and external constituencies” [28, p. 264]. But Hirschheim and Klein continue by noting that “of course a potentially thorny question is what are the IS sub-specializations and where would such a listing of them come from” [28, p. 263]? In this work, we consider two criteria for the identification of subfields. First, the body of work or set of researchers must be of such a size that it is recognizable and would likely have several journals dedicated to a set of topics of interest. Second, subfields are composed of interrelated topics that form a subject complex [3] or a network of topics with a cumulative tradition that become the focus of a community of researchers editors, reviewers, and journals over time. The subfields should be expected to surface from analyses conducted on the same kind of data with several years between analyses.

A fuller understanding of the subfields of IS also has implications for tenure processes and journal rankings. Lowry et al. [46] suggested that rankings of the journals of the whole IS field create problems for researchers that conduct their work exclusively within a sub-discipline. They went on to state that “by leaning toward ‘mainstream’ journals, sub-disciplines are inappropriately devalued, and their adherents are marginalized from the field of IS” [46, p. 36]. If researchers self-select into subfields, far-reaching implications for evaluation of research within the overall IS domain exist [10].

Given the uncertainty surrounding what actually constitutes subfields, this current research seeks to confirm the existence as well as the identity of the subfields which make up the IS discipline. With this information, future ranking studies as well as studies designed to determine the bodies of knowledge of subfields may provide a greatly increased understanding of the overall IS domain.

Our assertion is that the structure and organization of the IS field and its literature is a social phenomenon that evolves as information technologies, their applications, and their impacts are researched. The structure of the field is represented by intellectual communities that observe, describe, research, and theorize about certain phenomena, and then present the results of these processes to the editors and reviewers of journals that are accepting of those ideas, research methods, and evaluative criteria. Intellectual communities can be delineated by what Kuhn [39] describes as a “disciplinary matrix.” This matrix consists, in part, of the topics of interest, problems, and exemplars, and criteria for evaluation shared by the community. We suggest that this matrix is reflected in the editorial policies of journals, and that the research deemed appropriate by the reviewers and editors is one means of identifying communities. This idea is well described by Kuhn [39, p 131] who states:

Some of the principles deployed in my explanation of science are irreducibly sociological, at least at this time... Whatever scientific progress may be, we must account for it by examining the nature of the scientific group, discovering what it values, what it tolerates, and what it distains.

As IS is transdisciplinary in nature, the boundaries of these communities are not likely to have clear demarcations. Instead the groupings indicate large-scale social structures which organize the intellectual communities that comprise the IS domain. The challenge, then, becomes one of finding a research method that can explicate these social structures.

3. The Latent Categorization Method

Latent Categorization is a method that quantitatively represents the implicit semantic relationships and the content of a corpus of text. The value computed for each text artifact can then be subjected to a clustering algorithm, factor analysis, etc. LCM begins by treating each text artifact as a set of words without structure. Non-content bearing words that occur only once or almost always in a text artifact, (e.g. “of”, “the”, and “be”) are removed. Articles, prepositions, pronouns, and conjunctions, as well as common adjectives and adverbs may also be removed.

The remaining words are stemmed to avoid having multiple forms of a word represented in the analysis. Stemming converts a word to a related form, i.e. it “conflates” the word. For example, stemming the words “walk”, “walking”, and “walked” reduces all three to “walk.”

The technical details of the subsequent numeric transformation and statistical processing steps may be reviewed in Larsen and Monarchi [44]. In brief, the method creates a sparse matrix with the unique stems as rows, the artifacts as columns, and the number of occurrences of a specific stem in a specific abstract as the cell value. This matrix was weighted using TFIDF before it was submitted to a singular value decomposition (SVD), and the resulting right singular vectors are clustered using an agglomerative clustering approach. The main weakness of the approach is that interpretation of the resulting dendrogram is left up to the researcher; thereby leading to massive amounts of work and potential flaws of human interpretation [44].

This research develops an extension named Automatic Node Naming (ANN) that addresses this shortcoming in the standard LCM. After the decomposition and clustering, LCM as applied in this research consists of Word Impact Analysis (WIA), Automatic Node Naming (ANN), and Word Activation (WA). Word Impact Analysis (WIA) approximates the marginal impact that a specific term in a cluster has on the proximity of the cluster to other clusters. That is, how does the presence of the term affect the joining of two clusters to form a new cluster? (Actually, the analysis is performed from the perspective of what would happen if the term was removed from the cluster.) WIA allows the determination of the amount of movement in the position of the cluster centroid that would occur if the word was not present in the cluster. Technical details on WIA are provided in Appendix A.

The Automatic Node Namer (ANN) is a simple rule-based system that navigates the dendrogram using a tree-traversal order (see Appendix B for technical details). An extension to LCM [44], ANN skips the leaf nodes by considering only those nodes in the dendrogram which are clusters composed of more than one leaf.

In summary WIA and ANN automatically process the nodes of the dendrogram in a consistent manner. The outcome is a set of named clusters. These topics may individually be analyzed using the WA technique, helping during the development of a definition for each topic.

3.1. Word activation

While WIA and ANN have focused on representing the whole dataset, the concept of word activation (WA) is different. Word
activation starts with a map of the relationships between all words in the whole dataset. Using this structure and the distances between words, information about a specific cluster is superimposed on the structure to show which of the words were actually influential and related within the cluster under examination. The graphic would indicate the 20 word-stems that were closest in the n-dimensional space to the location of the computing cluster. However, while words that are close together have a high probability of appearing in the same cluster, many do not. This presents an opportunity to understand the specifics of what went into creating a cluster while also understanding what did not go into creating that same cluster.

The example shows the relationships of the cluster named "computing" with related words in the cluster. The thicker the line between words, the stronger the relationship is within this cluster. Items that are far apart, but with thick bonds indicate items that for the whole dataset are not close, but which within this particular cluster are quite related. Items without lines, while related to computing are not important within the context of this specific cluster.

Examining the topic of computing in the HCI community, we find that attitude and experience and reasonably close to anxious, age, gender, male, and female. It is also moderately close to adult, traits, and efficacy. Words that are close to computing for the whole dataset, but not close in this cluster, include literacy, like, technophobe, computer, microcomputer, aptitude, bout, and virus. This gives a good idea that this particular cluster is focused on attitudes of the process of computing and how experience and other variables, including anxiety, age, and gender, affect such attitudes. It is particularly instructive to observe the tight network of relationships between the activated words.

The cluster is not, however, focused on the hardware side as signified by the absence of links to computer and microcomputer, the literacy on such computers, computer aptitude, or technophobes. Neither is this particular cluster of research focused on computer viruses. It is fascinating to note that, in this example one might observe evidence suggesting stove-piping within the IS field. A reasonable researcher might observe that concepts such as technophobe, liking, literacy, and aptitude deserve some attention during computing attitude research. Taken together, these relationships provide information that may be used to define the individual clusters.

3.2. Sample selection and application of LCM

LCM provides the power to analyze a much larger and thereby potentially more representative journal dataset than has been previously examined. As noted earlier, exclusion of journals based upon geography (North American vs. European) or selection of only “mainstream” journals effectively pre-determines the research topics of the field because the selected subset of journals will reflect only certain intellectual communities. This selection of journals was based upon ranking of IS journals [e.g., 48, 59, 61]. An overview of the included journals may be found in Tables 3–9.

Abstracts from the selected IS journals were downloaded from automated databases, typed in manually, or retrieved through contacts with journal editors. To avoid disproportionate focus on journals with long publication histories over others, it was decided to include only articles published from 1990 through 2002. To allow further analysis of topics in individual journals, only journals where a minimum of 25 abstracts had been collected were included. One notable exclusion in the original analysis was Management Science, which publishes some IS papers every year. Because Management Science does not itself note which papers are considered IS, the journal was originally excluded to avoid creating a tautology by applying a subjective definition of what topics are considered IS. Only after a more advanced understanding of the IS field was attained through analysis of the pure IS journals were IS articles from Management Science selected for analysis (analyzing all abstracts would likely have led to communities created for reference disciplines). The journal was then added to the existing data, and the same analysis conducted to determine community membership for the journal. In the results section, the journal is listed with its community but noted because it was projected into the semantic space after-the-fact.

3.3. Factor analysis of cluster structure

To understand the intellectual organization of the IS field, we examined the relationships between journals and research topics by performing a factor analysis using the journals as variables and the clusters as observations. Each cell of the data contained the number of artifacts within a specific cluster that came from the specific journal.

Principal component analysis (PCA) with varimax rotation is the most common approach to factor analysis, and the orthogonal approach used in varimax fits with the LCM approach, which is also orthogonal. When the variables all have the same units, as they do here, the factors may be contrasted one versus another. The factor scores can therefore be used to understand which topics were most influential within each factor or community. The scores were also comparable across communities so that analysis of across cross-community topics was possible.

4. Results: structure of the IS field

4.1. Intellectual communities in information systems

Although the list of 324 research topics uncovered by this research is of interest, it alone does not provide satisfactory insight into the intellectual structure of the field. Such structure was revealed in the seven subfields or intellectual communities of journals and research topics presented below. The factor analysis (Table 2) accounted for 53% of the variance using seven factors as suggested by the Cattell scree test [9,29]. This is an acceptable level of variance for exploratory factor analysis according to contemporary standards [see, e.g., 41,61], especially when such a high proportion is derived from only 12% of the potential set of factors. While the Cattell test may sometimes result in an insufficient number of factors, all factor solutions between two and six were examined and all excluded a significant number of journals as unexplainable by that factor solution. The Kaiser–Meyer–Olkin Measure of Sampling
Adequacy of .77 was well above the threshold value of .50 [33,58]. Bartlett’s Test of Sphericity further shows significance below .001, well below the required cutoff point of .05, suggesting that factor analysis was appropriate for this dataset.

The following tables (Tables 3–9) consist of two groups of columns. First, the topic areas of each community of interest are split into four columns. The columns present the main topics, how many abstracts clustered into that topic, a definition for the topic, and factor scores. The order of the topics was determined using factor scores derived through regression analysis, where community scores for each topic were computed from the coefficients. Sorting the community scores illustrates which topics are important in each community. To save space, only topics with a factor score ≥ 3.0 are defined in the table. The last row of each table lists research topics with factor scores between 3.0 and 1.0.

The second group of columns outlines the journals around which researchers of the community focus their submissions.

### Table 2

Total variance explained

<table>
<thead>
<tr>
<th>Component</th>
<th>Rotation sums of squared loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>1</td>
<td>8.096</td>
</tr>
<tr>
<td>2</td>
<td>4.713</td>
</tr>
<tr>
<td>3</td>
<td>4.598</td>
</tr>
<tr>
<td>4</td>
<td>4.022</td>
</tr>
<tr>
<td>5</td>
<td>4.015</td>
</tr>
<tr>
<td>6</td>
<td>3.601</td>
</tr>
<tr>
<td>7</td>
<td>2.913</td>
</tr>
</tbody>
</table>

The factor loadings are the components of each factor in terms of the original variables. Journals with loadings ≥ .3 are listed in the right-hand column of the community tables below. Journals that loaded highly on more than one factor were listed in multiple communities with an asterisk when listed in their non-primary community location.

**Community 1 — Management Information Systems** research (Table 3), with Information and Management as the most representative journal, focuses on the organization and individuals within the organization, as well as the tools used to improve the efficiency of organizations. The topics as well as the journals indicate that this factor might arguably represent the main thrust of what is often referred to as Management Information Systems research. It is interesting to note that the most mainstream and highly ranked IS journals – such as *MIS Quarterly, Information Systems Research*, and *Journal of Management Information Systems* are found in this community.

This community is quite heterogeneous and accounts for almost twice as much variance as any other community in the analysis (13.5%). Intuitively it appears that this community should be subject to further subdivision. We therefore redid our analysis for only the articles in journals that loaded on factor one (MIS community) in the first analysis at a factor loading of .5 or higher (selecting .5 prevented selection of articles that primarily belonged in another factor). For these 14 journals (including Management Science), there was a clear unidimensional structure.

Our research agrees with some previous papers and combines a number of topics within this community. It is important, however, to distinguish between an intellectual

### Table 3

Management Information Systems (MIS) community

<table>
<thead>
<tr>
<th>Community topic areas</th>
<th>Community journals</th>
<th>Factor loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management 145</td>
<td>Management of businesses and organizations and information technology.</td>
<td>5.66 European Journal of Information Systems</td>
</tr>
<tr>
<td>Strategy 148</td>
<td>Business strategy and relationships to IS strategy.</td>
<td>5.01 MIS Quarterly</td>
</tr>
<tr>
<td>Planning 143</td>
<td>Planning for systems. Related to strategies, methodologies, and integration.</td>
<td>3.64 Journal of Information Technology Management</td>
</tr>
<tr>
<td>EDI 97</td>
<td>Research on Electronic Data Interchange (EDI) and its value.</td>
<td>3.23 Journal of Management Information Systems</td>
</tr>
<tr>
<td>Firms 92</td>
<td>Firms investments, and strategy.</td>
<td>3.2 The Journal of Strategic Information Systems</td>
</tr>
<tr>
<td>BPR 88</td>
<td>Business Process Re-engineering, Business Process Redesign. Processes and change.</td>
<td>3.05 Journal of Systems Management</td>
</tr>
<tr>
<td>Projects 145</td>
<td>Research on projects for software developments.</td>
<td>3.01 Journal of Information Technology</td>
</tr>
</tbody>
</table>


- Asterisk denotes one of 16 journals that loaded on two communities.
- Two asterisks indicate Management Science which was projected in after the other analysis.
community and a topic. Every intellectual community contains research on many topics, and each researcher will (with a few notable exceptions) work on several topics. Some researchers may also do cross-community research. Because of this, the approach of analyzing individual papers provides advantages over the previous quantitative approaches of examining individual researchers.

**Community 2 — Global and Societal research** (Table 4), is focused on the ways in which information technologies influence developing countries and society in general. The *Electronic Journal on Information Systems in Developing Countries* is the most representative journal, and topics unique to this community include countries, culture, internet, global, health, and social. The placement of the *Journal of the AIS* in this category is surprising, but this may be due to a small sample of abstracts from this journal (25 abstracts).

**Community 3 — Human–Computer Interaction research** (Table 5) focuses on the design, use and interaction of humans and computers. This community supports research on interfaces, visualization and computing at the individual and group levels. Other topics unique to this community include work, programs, text, tasks, language, and hypertext. The community is best represented by the journal *Behavior and Information Technology*.

**Community 4 — Electronic Commerce research** (Table 6) publishes articles on topics such as electronic markets and modern technologies such as the Internet, the World Wide Web, and virtual technologies. The community is best represented by the *International Journal of Electronic Commerce*.

**Community 5 — Systems and Software Engineering research** (Table 7) is focused on computer software testing and verification, architecture, formal specifications and other aspects of IS engineering. Within this community, *IEEE Transactions on Software Engineering* is the most representative journal.

**Community 6 — Information Storage and Retrieval research** (Table 8). Although it shares topics with Community 5 (algorithms, objects and parallel processes), this community is differentiated by its interest in reliability, architecture, real-time computing, and software reuse. Information retrieval is the focus, whether structured and database-related, or unstructured.

### Table 4
**Global and Societal (GS) research**

<table>
<thead>
<tr>
<th>Community topic areas</th>
<th>Community journals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topics</strong></td>
<td><strong>Factor score</strong></td>
</tr>
<tr>
<td>Countries</td>
<td>133</td>
</tr>
<tr>
<td>Culture</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Asterisk denotes one of 16 journals that loaded on two communities.

### Table 5
**Human–Computer Interaction (HCI) research**

<table>
<thead>
<tr>
<th>Community topic areas</th>
<th>Community journals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topics</strong></td>
<td><strong>Factor score</strong></td>
</tr>
<tr>
<td>Interfaces</td>
<td>190</td>
</tr>
<tr>
<td>Design</td>
<td>186</td>
</tr>
<tr>
<td>Visualization</td>
<td>120</td>
</tr>
<tr>
<td>Knowledge</td>
<td>360</td>
</tr>
<tr>
<td>Human interaction</td>
<td>97</td>
</tr>
<tr>
<td>Computing</td>
<td>107</td>
</tr>
<tr>
<td>Users</td>
<td>108</td>
</tr>
<tr>
<td>Students</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Asterisk denotes one of 16 journals that loaded on two communities.
and document-based. This community has *Information Systems* as its most representative journal.

Community 7 — Knowledge-Based Systems research (Table 9) is focused on intelligent systems. This community’s focus includes knowledge systems, agent-based system, expert systems, and several related topics. Many of the topics overlap with other communities, and the community is best represented by *International Journal of Intelligent Systems in Accounting, Finance and Management*.

4.2. Trans-community topics

An additional analysis of the topics and journals revealed topics that appeared in multiple communities. These topics are significant because they are of intellectual interest to a wider community of editors, reviewers and journals. This interest by the greater IS community may indicate that they are considered important, and possibly central, themes to IS research. Topics that appear in more than two communities include groups, projects, business, expert, models, knowledge, web, students, virtual, design, networks, and mobile.

4.3. Validation of framework

The starting point for the factor analysis leading to the final framework was the combination of journals and topics, so these building blocks were selected for the validation effort. Because of the inclusive nature of the study we did not expect to find experts who understood journals and topics for the whole IS field, and therefore designed the study to take advantage of the respondents’ areas of expertise. The respondents were evenly selected from the editorial boards of all the journals in our study, and 270 online questionnaire requests were sent out. 24 requests were returned as undeliverable or suggesting that the respondent would be out of the office until after our deadline for finishing the survey, leaving 140 useable responses for a response rate of 57% after one follow-up request.

The survey was designed to give each respondent an open-ended option of listing the three journals most similar to the focal journal they were contacted about. It was decided to make this open-ended to examine the representativeness of the selected journals. Of the 325 journals listed by the respondent, 180 or 55% were included in our study. Given that databases of journals in our field range between 376 (Oklahoma State database) and 576 (Deakin University database), this indicates that while our study included only 10–18% of the journals in these databases, the journals selected were on respondents’ minds.

Using the seven communities suggested by LCM, it was examined whether human experts selected as most similar those journals that belonged in the same community as the journal for which they were on the editorial board. With the

---

**Table 6**

Electronic Commerce (EC) research

<table>
<thead>
<tr>
<th>Community topic areas</th>
<th>Community journals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topics</strong></td>
<td><strong>#Abs</strong></td>
</tr>
<tr>
<td>Internet</td>
<td>135</td>
</tr>
<tr>
<td>Web</td>
<td>164</td>
</tr>
<tr>
<td>Virtual</td>
<td>109</td>
</tr>
<tr>
<td>Commerce</td>
<td>75</td>
</tr>
<tr>
<td><strong>Community also includes:</strong> Online, Communication, Job, Expert, End-users, Language, Cognition, Hypertext, Errors, Training, Virtual, Computing, Context.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 7**

Systems and Software Engineering (SSE) research

<table>
<thead>
<tr>
<th>Community topic areas</th>
<th>Community journals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topics</strong></td>
<td><strong>#Abs</strong></td>
</tr>
<tr>
<td>Programs</td>
<td>229</td>
</tr>
<tr>
<td>Test</td>
<td>152</td>
</tr>
<tr>
<td>Software</td>
<td>226</td>
</tr>
<tr>
<td>Verification</td>
<td>109</td>
</tr>
<tr>
<td>Algorithms</td>
<td>206</td>
</tr>
<tr>
<td>Architecture</td>
<td>77</td>
</tr>
<tr>
<td>Formal specifications</td>
<td>88</td>
</tr>
<tr>
<td><strong>Community also includes:</strong> Objects, Maintenance, Protocols, Real-time, Codes, Parallel, Components, Projects, Language, Graphs, Net, Metrics, Mobile, Scheduling, Design, Distributed, Security, Client/Server, Code, Tools, Trees, Networks, Reuse, Requirements, Methods, Logic.</td>
<td></td>
</tr>
</tbody>
</table>

Asterisk denotes one of 16 journals that loaded on two communities.
exception of one journal, the *Journal of the AIS*, the expert agreement with LCM was remarkably high. For JAIS, it was clear that almost all the respondents mentioning it or who were on its editorial board considered it part of the MIS community. This confirmed our earlier worry about this journal having just 25 abstracts in our analysis, and that its placement was strongly dependent on the small sample which was most likely not representative for later articles. After JAIS was removed from the study, an 89% agreement between LCM and experts was reached, with a Cohen’s kappa of .83, generally acknowledged as “almost perfect” [43].

The second part of the validation survey concerned whether the expert respondents would pick out the same topics found by LCM to belong to the community that LCM had determined. Given that neither the journal placement nor the topic placement could overlap with expert rater opinion unless they both made sense, we expected a good kappa here as well, though moderated for the much higher cognitive strain of this task. The raters were given 42 topic names and definitions and asked to select between two to five topics that best represented the focal journal. After removing the *Journal of the AIS*, a Cohen’s kappa of .62 was achieved. While this is generally considered a “substantial” level of inter-rater agreement [43], it was so much lower than the kappa for the journal exercise that it requires further discussion. One possibility is that while the underlying analysis was solid, the naming procedure did not properly represent the content of each cluster or topic. Another possibility is that the second task was cognitively more difficult in that it required reading much text and then comparing 42 topics. Finally, the respondents were also asked not simply to mention the topics they themselves generally dealt with, but rather the topics the journal had in common with other journals in that category, and because many of these journals have wide scopes individual reviewers may not have understood all aspects of a journal’s focus. In conclusion, it is clear that the kappas for both tasks are well beyond acceptable, suggesting that the results from LCM were valid.

5. Discussion

The structure and identity of the IS discipline has been a persistent question since the field’s inception. This research avoided many of the problems and limitations of prior studies to show that the IS discipline is composed of seven identifiable

---

### Table 8

Information Storage and Retrieval (ISR) research

<table>
<thead>
<tr>
<th>Community topic areas</th>
<th>Community journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topics</td>
<td>Factor score</td>
</tr>
<tr>
<td>#Abs</td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td></td>
</tr>
</tbody>
</table>

| Queries 142 | Research on database queries. | 8.51 | Information Systems | .801 |
| Databases 119 | Database research, including queries, and design. | 5.71 | Journal of Database Management | .706 |
| Objects 213 | Research on objects and object orientation, including programming and databases. | 5.59 | ACM Transactions on Information Systems | .681 |
| Algorithm 206 | Research on algorithms. | 5.06 | Computer Journal* | .595 |
| Models 191 | Research focusing on model development | 4.32 | ACM Computing Surveys | .589 |
| Protocols 173 | Research on protocols for transactions, especially in a database sense. | 3.85 | Data Base for Advances in Information Systems | .532 |
| Documents 91 | Research on documents, both document retrieval and IS development related documents. | 3.77 | Journal of the ACM | .477 |
| Community also includes: Retrieval, Web, Trees, Data, Graphs, Parallel, Temporal, Files, Hypermedia, Search, Groups, Hypertext, Images, Mobile, Ethics, Methods, Link, Memory, Multimedia. | | | | |

Asterisk denotes one of 16 journals that loaded on two communities.

---

### Table 9

Knowledge-Based Systems (KBS) research

<table>
<thead>
<tr>
<th>Community Topic Areas</th>
<th>Community journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topics</td>
<td>Factor score</td>
</tr>
<tr>
<td>#Abs</td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td></td>
</tr>
</tbody>
</table>

| Knowledge 360 | Knowledge research. Ties to ontologies, Knowledge-Based Systems (KBS), and experts. | 8.99 | International Journal of Intelligent Systems in Accounting, Finance & Management | .777 |
| Agents 149 | Research on agents. | 7.45 | Expert Systems with Applications | .752 |
| Audit 78 | Research on auditing of expert systems and on auditing expert systems. | 5.60 | International Journal of Human–Computer Studies* | .459 |
| Networks 225 | Neural networks, trees, and routing. | 4.25 | Autonomous Agents and Multi-Agent Systems Journal of Information Systems | .426 |
| | | | International Journal of Man–Machine Studies* | .361 |
| | | | Information Systems Frontiers* | .345 |
| | | | Journals loading below .3 were not included in the list. | .341 |


Asterisk denotes one of 16 journals that loaded on two communities.
intellectual communities or subfields. While this research demonstrates that the IS field has subfields with clear demarcations, the inclusion of some topics in multiple communities demonstrates porosity of boundaries that characterizes vibrant intellectual communities and indicates research topics which are of interest to the broader IS field. However, although these different communities work on some common topics, the existence of different communities with different publication outlets suggests that researchers in different communities may often be unaware of complementary or overlapping research interests.

By identifying the intellectual communities underlying the IS discipline, this study provides an answer to Hirschheim and Klein’s [28] “thorny question,” of where to begin defining the “body of knowledge” for sub-specialization in the field. They have argued that this is of tantamount importance before the discourse of the IS field can be lifted to a new level. By examining the latent characteristics of abstracts published in a sample that was an order of magnitude larger than anything analyzed before, the central topics and intellectual communities of the IS field were identified. This is the first step to determining the bodies of knowledge of the subfields and increasing communication among them.

Knowledge of an academic discipline’s identity provides researchers with the ability to communicate with internal and external constituencies alike. Because certain classes of problems cannot be solved without communication between subfields [36], increasingly insular subfields decrease the probability of successful research outcomes. Communication problems with external constituencies decrease the likelihood that academicians and practitioners can successfully find common ground. Because of the importance of both internal and external communication, the discourse related to finding and communicating the IS field’s identity takes on paramount importance.

Discourse about the IS field’s identity and composition has reemerged in the context of journals ranking and “crisis in the IS field”. This research has shown just how open a journal may have to be to constitute a true umbrella journal for the IS field, rather than just the MIS community. In this study, we find that the top ranked journals in IS, including MISQ, ISR, and JMIS have focused their publication activities primarily in the MIS research community.

When comparing previous studies of the structure of the field to the seven intellectual communities found in this study, some striking observations may be made. Of the seven previous studies examining journals (see Table 1), four cover only journals in the MIS Research community. Two studies cover only two communities: (MIS and Information Storage and Retrieval) and one study covered three communities (MIS, Information Storage and Retrieval, and Systems and Software Engineering). Surprisingly, four communities (Global and Societal, Human–Computer Interaction, Electronic Commerce, and Knowledge-Based Systems) are supported by journals that were not sampled in any of the prior studies. This is due in part to the newness of the communities relative to the timeframes of the prior studies. But it also illuminates the evolutionary nature of the field and the need to include a broad sample of journals if we want to understand the entire discipline, rather than only those parts represented by a few top journals.

The boundary porosity between communities is of special interest in this research. The MIS community shares topics with five other communities, indicating a large degree of boundary porosity. The topics web, internet, virtual and service are held in common between the Electronic Commerce and Global and Societal communities, suggesting interest in the influence and application of e-commerce on the societal and global levels. In contrast to the findings of Glass et al. [24] who found little overlap between the disciplines of IS and software engineering, there are many topics shared between the Information Storage and Retrieval and Software Engineering communities within the IS field. This difference in our research may be due to the larger number of journals and articles included in this current study, as well as the avoidance of existing frameworks for analysis and sorting.

This research provides the opportunity for a new type of journal ranking study based upon publications within the set of seven communities. Such rankings would arguably lead to fairer evaluations of and by IS researchers who conduct most or all of their work within one of the communities. Recognizing these communities may help researchers to target publications outlets for their research topics, and should help tenure and promotion committees evaluate researchers’ production.

This study also introduces Latent Categorization and a new extension as a new analytic method for analysis of the burgeoning volume of textual data available to IS researchers. Previously the sheer size of textual datasets that might provide interesting insights and theory development for IS phenomenon limited researchers’ ability to analyze the data. LCM also provides for replicable analyses by avoiding many of the categorization and interpretation problems to which many qualitative methods are prone.

By itself, LCM provides replicable clusters of text artifacts. Unfortunately, while such clusters may be used for further statistical analysis such as factor analysis, it is essentially uninterpretable except though overlaying human analysis. The need for such an overlay mitigates the advantage of being able to analyze large datasets. Further, the need to cut a dendrogram at a certain level without taking the underlying semantics into consideration introduces problems in interpretations and analysis. By applying word impact analysis, LCM is capable of analyzing truly large datasets (orders of magnitude larger than the one used in this paper). By adding ANN, a semantically natural cutoff point is selected that allows better understanding of the underlying content and clusters that have a higher latent semantic consistency.

This study adds to our understanding of the IS landscape by identifying the topics researchers and editors actually select for publication and how the field is organized by research outlet. To the extent that one agrees with the normative suggestions for restricting the “identity” of IS to the nomological net surrounding the information technology artifact, some of these communities will find it easier to do so given their topics of interest. But research in IS, as it has been practiced and published in the past, has multiple identities which will most likely continue to defy well-defined boundaries as technology itself evolves and finds new personal, social, and cultural niches to fill.

6. Limitations

This study is subject to several limitations including the removal of a few journals; the inability to categorize an abstract in more than one cluster; the retention of some words unrelated to the topics under consideration, and the limited presence of human judgment.
It is important to note that the low number of data points for some journals leads us to remove these journals from the sample and to question two journals' placement due to their short publication history. In addition, the aggregation of the abstracts over the thirteen-year period leads to a snapshot — rather than a history — of a dynamic and evolving field. Future work should take this evolution into account and show how new research areas and communities have developed over time. Further, during the factor analysis, three journals loaded at less than .3 on any factor, leading us to remove them from further analysis: Accounting, Management, and Information Technologies (currently Information and Organization), the Journal of End User Computing, and the Scandinavian Journal of Information Systems. While this is not uncommon during factor analysis, future work including more IS journals may find additional small intellectual communities containing these and other journals.

A significant limitation of this research was that we were not able to categorize an abstract in more than one cluster. Clearly, some articles deal with multiple topics, and should be placed in more than one cluster. This is an potentially fruitful area for future research.

Although the resulting 53% variance accounted for after the factor analysis was acceptable for exploratory analysis, it indicated that there may be other smaller communities in coexistence with the identified seven major communities of the field. Future research should identify these sub-communities and their relationships to the main communities. Such an analysis may find that the information systems development subfield examined by Hirschheim and Klein [28] represents such a small community.

Acknowledgements

This paper has benefited greatly from comments by Claude Banville, Henri Barki, Akhil Kumar, Detmar Straub, and Zoya A. Voronovich. We also owe a great debt of gratitude to editors of several of the journals covered by this analysis for going to great lengths to grant us access to data.

Appendix A. Word Impact Analysis

This section begins by looking at two leaves being joined together, and then extends the approach to include the joining of clusters.

Let \( i \) and \( j \) be the two leaves being joined to form a new cluster. Choose one of them, say \( j \), and copy the \( j \)th column of the weighted term-frequency matrix, \( \mathbf{A} \), to a new vector \( \mathbf{q}_j \), calling it a pseudo-artifact (a pseudo-document in Deerwester, 1990).

In the \( k \)-dimensional space of \( \mathbf{A}_k \), the pseudo-artifact may be written as

\[
\mathbf{q}_j = \mathbf{A}_j^T \mathbf{U}_k \mathbf{S}_k^{-1} \tag{0.1}
\]

[8]. The weighted artifacts with squared Euclidian distance as the dissimilarity measure can be written as

\[
\mathbf{A}_k^T \mathbf{A}_k = \mathbf{V}_k \mathbf{S}_k^2 \mathbf{V}_k^T \tag{0.2}
\]

Using Eq. (0.1) for the approximation to artifact \( j \) and Eq. (0.2) for artifact \( i \), the cosine between them is

\[
\cos \theta_{ij} = \frac{(\mathbf{v}_j^T \mathbf{S})(\mathbf{U}^j \mathbf{q}_j)}{||\mathbf{v}_j^T \mathbf{S}||_2 \|\mathbf{U}^j \mathbf{q}_j\|_2} \tag{0.3}
\]

where the subscript \( k \) has been dropped for clarity. (This is also true in all subsequent equations.)

In turn, because \( \mathbf{U} \), \( \mathbf{S} \), and \( \mathbf{V} \) are all known, Eq. (0.3) can be written as

\[
\cos \theta_{ij} = c_0 \sqrt{\sum_{h=1}^k (\mathbf{V}_h \mathbf{S}_h)^2} - \sqrt{\sum_{h=1}^k \mathbf{U}_h \mathbf{q}_h)^2} \tag{0.4}
\]

where

\[
c_0 = \sqrt{\sum_{h=1}^k (\mathbf{V}_h \mathbf{S}_h)^2} \tag{0.5}
\]

and \( \mathbf{e}_j^T \) is the \( j \)th column of a \( d \times d \) identity matrix.

Eq. (0.4), can be used to determine the difference that the presence or absence of a specific term in the query, \( q_h \), has upon \( \cos \theta_{ij} \). The approach is simply to compute the cosine with and without the term in the query, \( \cos \theta_{ij} \) and \( \cos \theta_{ih} \), respectively. The difference between them divided by \( \cos \theta_{ij} \) and then subtracted from 1 provides a measure of the impact of this term on the association between the \( i \) and \( j \) artifacts. (Note that the sum of these proportions will not, in general, equal 1 because of the interactions of the terms in the denominator of Eq. (0.4).) The result is labeled \( \rho_{ih} \).

\[
\rho_{ih} = (\cos \theta_{ij} - \cos \theta_{ih})/\cos \theta_{ij} \tag{0.6}
\]

with larger values indicating greater impact.

The procedure is to display the \( \rho_{ih} \) in decreasing magnitude to determine the principle determinants of the value of the cosine, \( \cos \theta_{ij} \). Experience suggests that at most 10 terms are the determinants for each cosine. (Note that the matrix of these cosines is not symmetric due to the interactions within the term–term matrix [for a discussion of the transitive effects of terms in LSI, see 37]). At this point, a knowledge structure may be constructed by examining this data for each point in the dendrogram.

Appendix B. Automatic Node Namer

As the Automatic Node Namer (ANN) navigates the tree, it goes through the following steps to develop the name for each cluster. In the description below, all points on the dendrogram are “nodes”. Leaves are the nodes at the bottom of the tree, the original data items which are being clustered. Clusters are groupings of two nodes which may or may not be clusters themselves. (Leaves cannot be clusters.) The two nodes which are members of a cluster may themselves be clusters.
1. For each pair of nodes which combine to form a new cluster, ANN first determines the list of words in each node whose contribution to forming the cluster exceeds a threshold set by the analyst.

2. ANN next combines the two lists into a proposed name based on whether the analyst has required that all candidate words for the name exist in both lists or in only one. Within the list, the words are in decreasing magnitude of their contribution (see step 1).

3. Then ANN invokes the following rule set.
   a. If either of the nodes is a leaf node, then the proposed name becomes the name of the cluster.
   b. If neither node is a leaf node, then
      i. if neither node has a name (i.e., each node's name is a zero-length string, ""), then the name of the cluster is set to a zero-length string also. (This rule is analyst-selectable.)
      ii. if only one of the nodes has a name (the other is a zero-length string), and all of the words in that name exist in the proposed name for this cluster, then the proposed name becomes the name. Otherwise the name for this cluster is set to ""
      iii. if both of the nodes have names, then the cluster name is set to the union of the words in the names of the two nodes and the words in the proposed name.

4. ANN moves to the next node by tree-traversal order until all nodes are examined and nodes fitting the criteria are labeled.

In this case, the dendrogram was automatically analyzed with a cosine cutoff of .10, a value that when tested worked well for small and large datasets alike. This left most of the dendrogram labeled with the words that best represented the well for small and large datasets alike. This left most of the words for the name in only one. The analyst was able to test the words that were selected.

References

Kai R. Larsen is an Associate Professor of Information Systems at the University of Colorado/Boulder Leeds School of Business. His research interests revolve around social influence in adoption of information systems and automatic knowledge extraction from text. He has published in several journals, including Communications of the ACM, Journal of Management Information Systems, Sociological Methodology, and the European Journal of Information Systems.

Doctor Monarchi is a retired Full Professor of Systems at the University of Colorado at Boulder (UCB), where he was a member of the faculty from 1972 to 2006. He has published in various academic journals, such as the Journal of the American Statistical Association, Communications of the ACM, and Decision Support Systems, and presented his work both nationally and internationally. Doctor Monarchi’s primary areas of interest are natural language programming, data and text mining, and knowledge extraction.

Dirk S. Hovorka is a Scholar in Residence at the Leeds School of Business, University of Colorado at Boulder. He received his B.A. from Williams College, MA and was awarded M.S. degrees in Geology and in Interdisciplinary Telecommunications as well as a Ph.D. in Information Systems from the University of Colorado. His research includes influences of social networks on knowledge exchange, the philosophical foundations of IS research, information technology tailoring, and information systems in science.

Christopher N. Bailey is a graduate of the Information Systems program at the Leeds School of Business, University of Colorado. He currently works for Wall Street on Demand.