Modeling Task-Genre Relationships for IR in the Workplace

Luanne Freund
Faculty of Information Studies
University of Toronto
Toronto, Canada
freund@fis.utoronto.ca

Elaine G. Toms
Faculty of Management
Dalhousie University
Halifax, Canada
etoms@dal.ca

Charles L.A.Clarke
School of Computer Science
University of Waterloo
Waterloo, Canada
clacarke@plg.uwaterloo.ca

ABSTRACT
Context influences the search process, but to date research has not definitively identified which aspects of context are the most influential for information retrieval, and thus are worthy of integration in today's retrieval systems. In this research, we isolated for examination two aspects of context: task and document genre and examined the relationship between them within a software engineering work domain. In this domain, the nature of the task has an impact on decisions of relevance and usefulness, and the document collection contains a distinctive set of genre. Our data set was a document repository created and used by our target population. The document surrogates were meta-tagged by purpose and document type. Correspondence analysis of this categorical data identified some specific relationships between genres and tasks, as well as four broad dimensions of variability underlying these relationships. These results have the potential to inform the design of a contextual retrieval system by refining search results for this domain.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval: information filtering, selection process

General Terms
Measurement, Design

Keywords
contextual information retrieval, correspondence analysis, work tasks, genre, enterprise search

1. INTRODUCTION
Context plays a powerful role in shaping how people search for information and in determining what information they select and use [1-3]. By making use of context, retrieval systems will be able to filter and/or rank information with a higher degree of specificity than is possible using the traditional 'bag of words' approach. The capability to provide targeted search results becomes increasingly important as the amount of information available within organizations and publicly on the Internet is growing so rapidly as to be overwhelming to the average searcher. Thus, it is not surprising that contextual search has recently risen to the top of the research agenda for information retrieval [4].

However, context is a very broad concept, and researchers in this field have not yet identified a concrete set of contextual factors that influence search behaviour, nor have they identified how these factors influence search. Our research is part of a larger study that is developing an approach to contextual search in the workplace. As part of the larger study, we conducted interviews with the target population – software engineering consultants – to understand which broad contextual factors influence how they search for and select information [5]. Emanating from that analysis were a discrete set of contextual variables, the most significant of which were *work tasks* and *information tasks* of searchers and *genres* - document types based on similarity of form and purpose. In this paper, we report on the next stage of the project: a focused analysis of these contextual factors. We hypothesized that a relationship exists between searchers' tasks and document genres. We believe that such a relationship can be exploited by an IR system to contribute to more precise results. Our goal was to test the relationship between these variables and to explore and identify more specific relationships and patterns of association between these variables.

2. PRIOR WORK

2.1 Enterprise Search
One area where contextual search has the potential to make a major contribution is in improving workplace search systems, or *enterprise search*. Enterprise search is a technical challenge due to the range of unstructured and structured data types and storage mechanisms in use, the paucity of hyperlinks, and complex security considerations within organizations [6, 7]. Due to the proprietary nature of most organizational information spaces, it has been difficult for researchers to gain access to large document collections in order to conduct research, so there is relatively little research data available. However, it seems that the smaller size and well-defined topic and task domains of organizational information spaces, as compared with the Internet, make them particularly suited to implementations of contextual search.

SIGIR Conference'05, August 15-1–2, 2004, Salvador, Brazil

2.2 Task and Genre

Within the workplace, there are two marked elements of context for information seeking and use: tasks and document genres. Peoples' information needs are typically inspired by their work tasks. A single work task may provide the impetus for any number of information searching and retrieval tasks, and thus forms a natural unit of context for workplace searching [8, 9]. On the information side, genre is a dominant feature of information resources in the workplace. Organizational theory views genre as a powerful carrier of pragmatic meaning. Genres are described as, "typified communicative actions characterized by similar substance and form and taken in response to recurrent situations" [10]. Genre repertoires, which are sets of commonly used document types, exist within work domains, and provide an organizing structure for information sharing [11]. Genre, like task, seems to be a key unit of context for information use within specific work domains. There has been some work on using categorization of document genres to support search [12-14] but these approaches do not consider how genres are related to the search context or information need. Some early steps in this direction are being taken in the context of the TREC HARD track [15, 16], but so far with a very limited set of genres.

Our study can contribute to this prior work because it looks at associations between specific tasks and genres, rather than at one or the other in isolation. The approach adopted in this study is based on the idea, that while it is clear that people have difficulty expressing information needs [2], it is reasonable to assume that in a work context, they will have little difficulty identifying the task they are trying to complete.

2.3 Software Engineers

There are a limited number of studies of the information behaviour of software engineers, but there is a broad literature relating to how engineers in general find and use information. Information plays a major role in the work of engineers, who spend about 40-66% of their time gathering and producing information related to their work [17]. In general, engineers and scientists need information in order to solve problems and complete tasks, and are often satisfied with a small amount of "good enough" information. Studies of engineers have found that they tend to rely heavily upon verbal communication with colleagues and internal reports as information sources [18-20]. There are a whole range of non-topical factors that influence engineers' selection of sources (people and documents). These include familiarity, the amount of time needed to use the source, the level of detail, and physical accessibility [21]. For software engineers more specifically, important features are: content, currency, availability, and the use of examples. [22]. Although information use among engineers seems to be closely tied to specific task contexts, studies of engineers to date have not closely examined the nature of the connection between tasks and information selection and use.

3. METHODOLOGY

3.1 Target Work Domain

We conducted our research in a large high-tech corporation working with software engineering consultants who provide clients with a range of services with respect to specific software products. The scope of their work and the knowledge required is very broad, so they rely heavily upon digital information, in keeping with the general profile of engineers as information consumers. The documents used by this group are widely dispersed on the company intranet, internal databases, and on external web sites. The documents exist in a range of genres, some more general, such as tutorials and presentations, and others that are specific to this domain, such as engagement summaries and integrated scenarios. Genre is used as a means of categorizing documents in most of the key repositories used by this group, although the genre taxonomies are not standardized. See [5] for an in depth description of this work domain.

3.2 Data

With the goal of identifying relationships between tasks and document genres in this domain we conducted an analysis of one of the key information sources for this group: an intellectual capital (IC) repository of documents either recommended or authored by the consultants. This "found" dataset is a restricted access Lotus Notes database that contains thousands of documents ranging from brief notes to lengthy review articles. The process of submitting items to the IC repository includes completing a metadata form with 30 data fields. Alongside the more common factual fields such as *author*, *date submitted*, and *title*, are two interpretive fields, *artifact type* and *purpose*, which we used for this analysis.

Artifact type identifies the document type or genre of an item, based on some characterization of its form and/or purpose. In most cases, a single artifact type value was assigned to each document.

Purpose identifies the end for which an item was to be used, which essentially represents the task/s that a document was meant to support. More than one purpose tags were frequently assigned to single documents.

. **Table 1: Categories for Artifact Type and Purpose Variables**

Artifact Types /Genres (17)	**Purpose: Work Tasks (20)**	**Purpose: Information tasks (16)**
architecture/design	administrate/install	compare
collection	architecture/design	contacts
cookbook	capacity planning	demonstrate
demo	competitive evaluation	document
discussion	configuration	educate
engagement summary	debugging	example/reuse
lecture/lab	deployment	guide/manual
legal material	development	index
presentation	discovery session	market/sell
product feedback	implementation	methodology
reading material	installation	reference
sales kit	integration	roadmap
schedule	migration	standards
source code	performance tuning	support
tools	proof of concept	technical info
website/repository	product presentation	tool
	project management	
	project review	
	security	
	test	

The underlying idea of our analysis is that the frequency of co-occurrence of certain genres with certain purposes, or tasks, is indicative of patterns of association between these variables.

The initial data set consisted of 6400 pairs of *purpose* and *artifact type* tags relating to items in the repository. However, despite apparent attempts to control data entry for each of these fields to a set of about 16 controlled values, there were actually 47 different artifact types and 125 different purposes in evidence in the dataset. Prior to conducting the analysis, we reduced this to a more concise set of 17 *artifact types* and 36 *purposes* (Table 1). We did this by grouping sub-topics together under more general headings; for instance *presentation/lecture* was included under presentation; and by removing values with very low frequency, such as the one *install tip.* The final reduced dataset contained about 5800 pairs of tags. In order to avoid introducing excessive bias, we retained the category labels assigned by the consultants, to the extent possible. This results in some lack of clarity in the data when tasks and genres are very closely associated with one another, such as the *tool/s* and *architecture/design* categories which are both genres and tasks.

On further inspection, we realized that the *purpose* tag had been interpreted in two ways; about half the values indicate work tasks and the other half indicate informational tasks. We divided these into two conceptual groups (see Table 1) for purposes of interpreting the results, but the analysis treated all 36 purposes as categories of a single variable This taxonomy of tasks and genres was developed by the designers and users of this repository, rather than by the research team, and as such is not meant to serve as a generic taxonomy for the software engineering domain, but does seem to reflect the environment in which this group operates.

3.3 Correspondence Analysis

There are relatively few tools available to analyze this type of data: two 'corresponding' or inter-related variables each with multiple unordered categories. The simplest form of analysis is a cross tabulation, or contingency table. In our case, the contingency table was 17 x 36, too unwieldy to support interpretation of the data in any significant way. For this we turned to correspondence analysis, a technique developed to identify and visualize associations between two or more categorical variables. Correspondence analysis calculates the associations between and among row and column values in two-way contingency tables by calculating a measure of distance between points (categories). It uses a Chi-Square distribution to determine the independence of cells and the similarity/dissimilarity of row and column variables, and represents them graphically in a low dimensional space [23]. The goals of correspondence analysis are similar to other multivariate analyses such as factor analysis and principle components analysis which are used for dimension reduction. However, where factor analysis determines which *variables* cluster together, correspondence analysis identifies clusters of *categories.* Correspondence analysis has been used in a wide range of fields, including education [24], marketing [25], and image retrieval [26]. Similar methods exist under a number of different names, including optimal scaling, dual scaling and homogeneity analysis [27].

4. DATA ANALYSIS & RESULTS

4.1 Analysis

Correspondence analysis was conducted using the Categories Module of SPSS 13.0. The analysis used the full matrix of 17 artifact types and 36 purposes in a 5800 item data set. The number of items in each genre category varied widely, from 9 (white papers) to 1629 (reading material), and similarly for the purpose categories – from 4 (contacts) to 1903 (educate). We used symmetrical normalization since we were interested in the general distribution of data-points and the relationships between row and column points. Using this approach means that although proximity between row and column variables indicates association, the distances between them cannot be interpreted as measured strengths of association.

The results supported our hypothesis that a significant relationship exists between task and genre (X^2= 5878.968, df=612, p<.001). The analysis calculates the proportion of each genre category that is associated with each task in the dataset. While this table is too large to include, some of its results are notable. For example, ~33% of all cookbooks (documents with step-by-step instructions) occurred together with only 2 tasks: *educate* and *administer/install*; the remaining 67% of cookbooks were spread thinly over a further 28 tasks, and 6 tasks had no cookbooks associated with them. In another case, ~80% of newsgroup discussion threads were associated with 4 different tasks: *competitive evaluations*, *educate*, *demonstrate* and *document.* The remaining 20% were distributed over only 12 tasks, and fully 20 tasks were completely free of association with discussion threads.

This sort of variability was present throughout the data set and can be used to identify pairs of genres and tasks that are positively or negatively associated in a general way. However, correspondence analysis allows us to go further by decomposing the variation between and among tasks and genres into separate dimensions that represent different sources or conditions of variability. The challenge is in being able to assign a 'class' concept to the dimension. Lower level dimensions, in particular, which account for little of the variance are often difficult to interpret. For this next step, we needed to determine how many dimensions to consider.

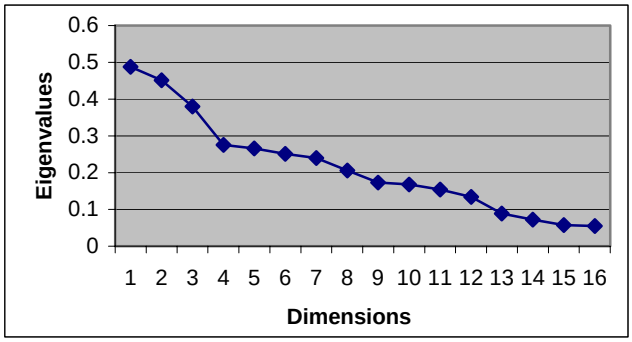
Table 2: Dimensions in the Dataset

Dimension	Singular Value	Inertia	Proportion of Inertia Accounted for
1	0.49	0.24	0.23
2	0.45	0.20	0.20
3	0.38	0.14	0.14
4	0.28	0.08	0.07
5	0.27	0.07	0.07
6	0.25	0.06	0.06
Total		.79	0.77

The maximum possible number of dimensions (16) accounts for 100% of the total variation or *inertia* (percent of variance explained by each dimension*)*. In order to be able to interpret the data, we considered the fewest number of dimensions that accounted for most of the variance. As illustrated in Table 2, the

first six dimensions accounted for 77% of the variance, with the remaining 10 dimensions accounting for the rest. Typically one to three dimensions account for this much variance, but our result is comparable to similar analyses of large matrices [24]. Table 2 also shows the correlation (analogous to Pearson correlation coefficient) between the row and column scores for each dimension (singular value score); the greater the inertia, the greater the association between row and column. Plotting the singular values in a Scree plot (Figure 1) indicates that after the 4rth dimension, the values begin to drop less rapidly and are similar in size. Thus, the Scree elbow test indicates that a 4 dimensional solution will be suitable for this dataset.

Figure 1: Scree Plot of Singular Values



4.2 Interpretation of Graphic Displays

The graphical output of our 4 dimensional solution consists of 3 correspondence maps, which are biplots (Figures 2-4) showing the first and strongest dimension (along the x-axis) by each of the remaining three dimensions (along the y-axis). Some labels are hidden for legibility. Interpretation of the broad distribution of the data points in these plots is aimed at understanding the nature of the underlying variables that influence the associations between the categories. In interpreting these graphs, it is important to recognize that not all points contribute equally to the inertia or variation of each dimension. Determining the relative contributions of different points supports more accurate interpretation. This can be done by calculating the average contribution of row and column points per dimension (1/ number of points) and comparing this with the contribution to inertia for each point in each dimension as reported in the output. A point making a higher than average contribution, is playing a more dominant role in determining the variation in that dimension.

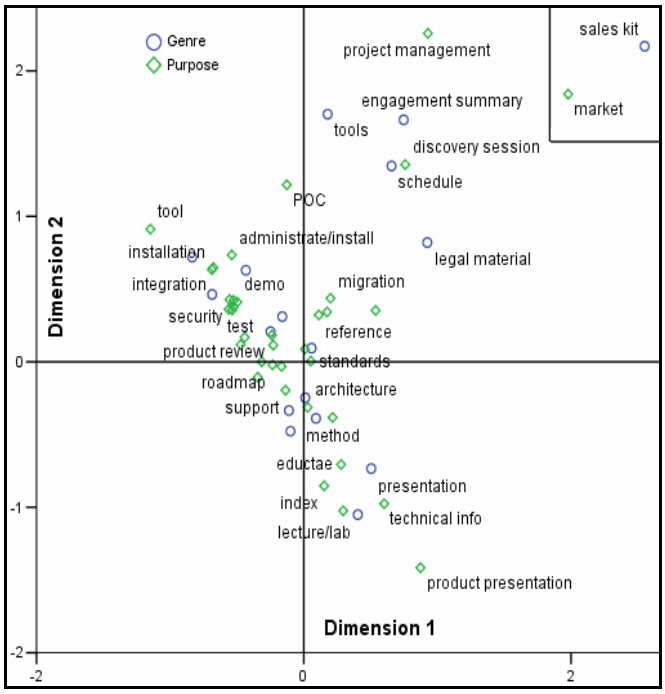
4.2.1 Dimension 1: Variability based on Work Activities

The first dimension of variability in associations between genres and tasks in this domain is represented by the horizontal distribution of points along the x-axis in Figures 2, 3 and 4. Our interpretation of the underlying cause of this variation is that there exists a strong distinction between the two kinds of activities in which software consultants are involved. The left quadrants of figures 2-4 include points related to software engineering: *development, deployment, debugging, source code, demo, test, guide*, etc. Points in the right quadrants are related to consulting activities: *project management, schedule, legal materials, product presentation*, etc. Looking at the data points with above average contributions to inertia supports this interpretation. In this dimension, the stronger points are clearly associated with either software engineering or consulting. Some strong genre points in this dimension (> .058 contribution to inertia) are *source code* (.123) and *sales k*it (.613) and some strong work task points (> .028) are *administrate* (.034), *integration* (.035) and *demonstrate* (.042). Points near the centroid are less strongly associated with the variance in this dimension, so we can find some cross-over areas along the x-axis such as *proof of concept, capacity planning* and *architecture*, which clearly require technical knowledge, but tend to involve more interaction with the customer. There is another cluster of two points - *marketing* and *sales kits* - far in the upper right quadrant that represent yet another work activity: sales. These can be considered outliers in this data, as they are clearly related to one another, but do not have a strong connection with the rest of the materials in this repository. This is not surprising, as the consultants in this group do very little sales work.

4.2.2 Dimension 2: Variability based on Information Goals: Doing vs. Learning.

The second dimension of variability in associations between genres and tasks in this domain is represented by the vertical distribution of points along the y-axis in Figure 2. By comparing the upper and lower quadrants of Figure 2, we interpreted this dimension as the distinction between a procedural, task-based and a knowledge-based perspective. Software engineering and consulting tasks as well as task–oriented and procedural information types appear in the upper quadrants. Some of the stronger points are *cookbooks* (.108) and *source code* (.099) on the software engineering side, and *engagement summaries* (.134) on the consulting side.

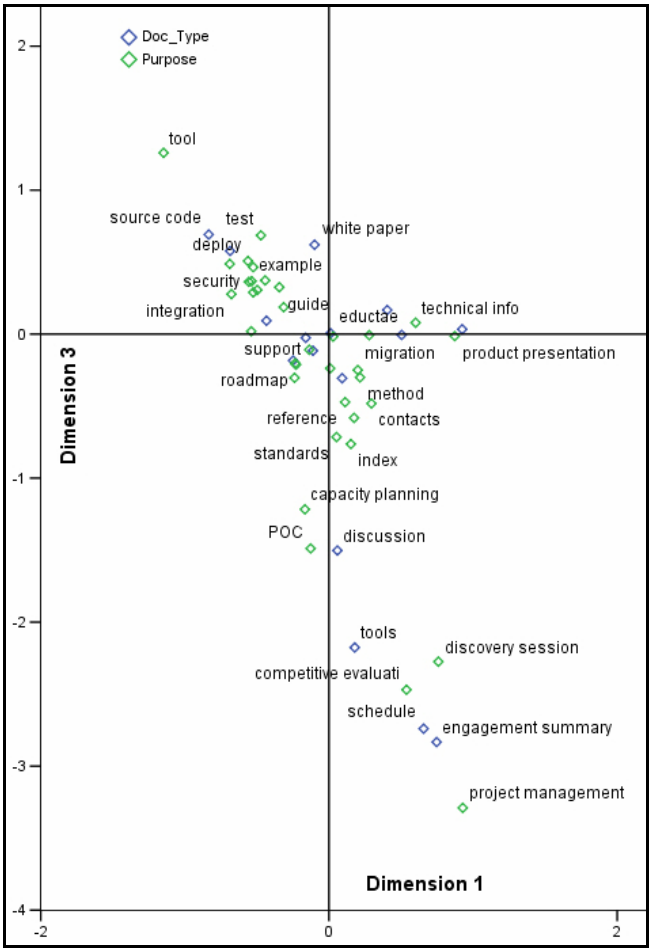
Figure 2: Correspondence Map of Dimensions 1&2



*datapoints in upper right were shifted to appear on the graph

In the lower quadrants, points represent more generic sources of information and learning as well as some higher level, knowledge-based tasks, such as *architecture* and *capacity planning*. The strongest points in the lower quadrants are all related to teaching and learning: *educate* (.363), presentation (.227) and *lecture/lab* (.096). There are no strong points associated with the lower left quadrant which represents high level knowledge/software engineering. The few data points that appear, however, do support the interpretation: *guide*, *website* and *white paper*.

Figure 3: Correspondence Map of Dimensions 1&3



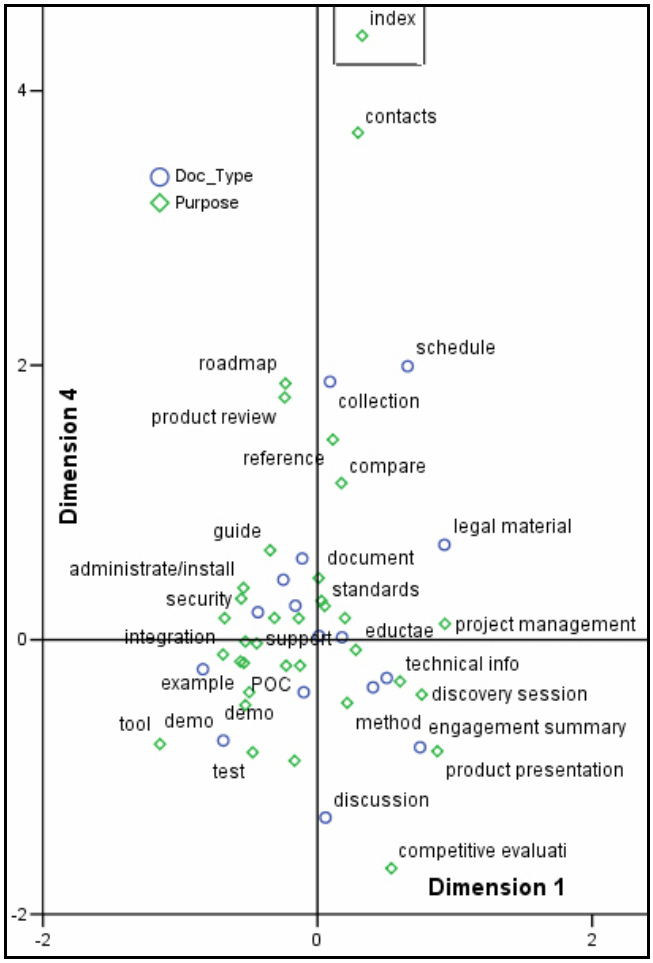
4.2.3 Dimension 3: Variability based on Tasks and Roles

The third dimension of variability is represented by the vertical distribution of points along the y-axis in Figure 3. There is a strong correlation between the first and third dimensions, which can be seen by the shape of the data (Figure 3). Most of the points are in the upper left (software engineering) and lower right (consulting) quadrants. However, the strongest points in the dimension point to a sharper distinction between various project roles which consultants in this group typically perform: project manager, architect/designer and technical specialist. The strong points *project management* (.251) and *engagement summary* (.460) are closely associated with the first role, and related points form a clear cluster in the lower right. There is another loose cluster just below the centroid, which seems to relate to the second role, as it includes a number of high level technical tasks and related information types: *architecture, design docs, capacity planning, standards,* and *roadmap*. Finally, on the far upper left is another cluster that includes most of the software engineering tasks, but the strongest points relate to more low-level technical tasks: *source code* (.108), *tool* (.058) and *demo* (.052).

Dimension 4: Variability based on Information Goals: Fact Finding vs. Demonstrating

The fourth dimension of variability is represented by the vertical distribution of points along the y-axis in Figure 4. Our interpretation of the underlying source of variation in this dimension is based on a distinction between use of reference materials to look up information and more interactive forms of information transfer, such as presentations and demonstrations.

Figure 4: Correspondence Map of Dimensions 1&4



*datapoint at upper edge was shifted down to appear on the graph

In contrast to dimension 3, there is very little correlation here with the distinction between software engineering and consulting: the

data points are spread evenly over the four quadrants. The upper quadrant relates to documentation and tasks that are related to reference materials. The strongest point in the upper quadrants is *index* (.331), and other related points are *reference*, *roadmap*, and *document*. Associated tasks are: *administrate*, *implementation* and *architecture*. This is contrasted with the points in the lower quadrant which reflect a more active use of information for teaching or demonstrating something*: demonstrate, test, presentation, lecture/lab, and discussion*. Strong points in the lower quadrant are *document* (.044), *test* (.024) and *competitive evaluation* (.149). Other tasks in the lower quadrant are *performance tuning* and *debugging* on the software engineering side, and *product presentation* and *discovery session* on the consulting side.

Table 3: Patterns of Association

	Software Engineering	Consulting
Doing **"how-to"**	**tasks** administrate integration installation debugging security configuration deployment development performance tuning **document types** tool cookbook roadmap demo	**tasks** project management discovery session **document types** tools engagement summary schedule legal material
Learning **"why"**	**tasks** architecture capacity planning **document types** guide website	**tasks** product presentation **document types** collection presentation technical info lecture/lab
Fact Finding **"what"**	**tasks** administrate security implementation **document types** website design docs roadmap standards	**tasks** project management **document types** index schedule legal material
Demonstrating **"show me"**	**tasks** test debugging performance tuning installation deployment **document types** source code tool demo	**tasks** discovery session product presentation competitive evaluation **document types** methods discussion engagement summary technical info

4.2.4 Summary of Dimensions

Table 3 provides a summary of the patterns of association identified in the different dimensions together with some of the more dominant tasks and genres for each. The strongest associations between tasks and document types are based on meta-task categories: software engineering and consulting. The other main associations are based on information goals: doing, learning, using documentation, and demonstrating.

4.2.5 Data Clusters

In addition to the broad patterns of association identified through analysis of the correspondence maps, there are some smaller groupings of points that tend to be strongly associated throughout the various dimensions. By plotting the row points (genres) separately (Figure 5), a number of these small groupings become more apparent. Interpretation of these clusters suggests some micro patterns of association within the data, a number of which are highlighted on the graph.

Reusables (*tools, engagement summary, schedule, legal material*). This cluster relates to consulting work. They are all examples of genres that would be used primarily as templates and reusable information objects for successive consulting projects. These materials are closely associated with the task *project management*.

Low-level technical (*source code, cookbook, demo*). These materials all contain low-level - concrete and specific - technical information. This type of information is closely associated with a grouping of procedural tasks with a similar low-level profile: *installation, configuration, deployment, administrate.*

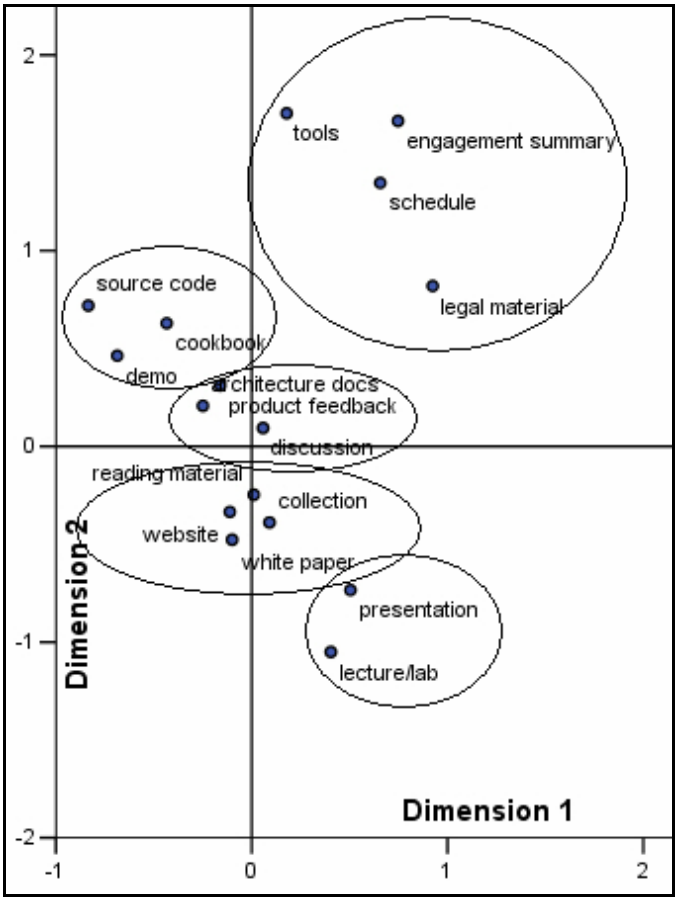
System data (*source code, demo, tool, test*). This cluster relates to system-generated information rather than secondary documentation and seems to be related to *debugging* and *performance tuning*. In some dimensions, this cluster appears together with more development-oriented software engineering tasks (as in Figure 5), but in others it appears as a distinct cluster.

Product maintenance (*product feedback, design documents, discussion threads*). This cluster has a product focus, and tends to straddle the axis between software engineering and consulting. Related tasks are *migration* and *support*.

High-level generic (*reading material, website, white paper, collection*). This cluster includes materials that consultants would use to find conceptual rather than procedural information about technologies and products. Associated tasks are high-level and few: *architecture* and *capacity planning*.

Educational (presentation, lecture/lab). These materials are closely related to one another and to the information task *educate*, but they do not have a strong associations with any particular work task or with either software engineering or consulting.

Figure 5: Genre Clusters



5. DISCUSSION

5.1 The Case for Contextual Factors

Using an existing data set that contained two related variables, document genre and purpose/task, we used correspondence analysis to explore that relationship. There are two key findings from this analysis.

First, the analysis indicated a strong association between the two variables, as well as identifying which genres tend to co-occur with which tasks, and notably, which do not. This suggests that in this domain, these are important contextual factors of information use, and that exploitation of these relationships could contribute to more effective work- and task-based information retrieval systems.

Secondly, the analysis identified a number of macro and micro patterns of association between and among these variables. Different work activities (software engineering, consulting, sales) seem to be a major source of variation among tasks and genres, as are different information tasks (doing, learning, fact finding and demonstrating). The patterns of association represented by the data clusters suggest that there are smaller and more subtle relationships based on different characteristics of tasks and information, such as conceptual level, a focus on products or technology, human or system generated information, etc. This suggests a complex relationship that might best be represented in an IR system through a multidimensional framework, in which, at a minimum, the searcher's problem = topic + work task + information goal, and the document = topic + genre.

5.2 Limitations

The data used in this analysis was created by a user community to support their information use environment. Thus, it suffers from the usual problems of real-world data – a certain amount of bias. Thus these metadata elements cannot be considered representative of all software engineers. The consultants who contribute to this particular data repository are technical experts, who consider it a good source of technical data; however, materials related more to products and consulting services are not concentrated in this collection. Perhaps this served to emphasize the distinction between software engineering and consulting tasks, at the expense of differentiating among software engineering tasks. Overall, the consulting tasks were better described by the 4-dimensional analysis than were the software engineering tasks, based on the reported dimension contribution scores.

The main challenge we faced concerned the nature of the data: nominal variables with multiple elements. This is not the typical raw material for statistical analyses, and does not support the calculation of correlation coefficients for genres and tasks. We were able to identify positive and negative associations, but cannot claim statistical significance in the classic sense. That said, we found correspondence analysis to be the best means possible to analyze this very rich user-defined data set.

5.3 Implications for IR

The key results of this analysis with respect to IR was in definitively associating genre with task and identifying a means of mining existing data to extract this relationship within a specific domain. The existence of a task – genre relationship suggests a means of supplementing traditional topical ranking algorithms by weighting document genres with respect to searchers' tasks. This approach has the potential to improve the situational relevance or *usefulness* of search results, particularly in an enterprise search context. Furthermore, unlike some IR systems that provide genre categories for filtering of the results, this approach does not require that the user know which genres they want to find, but only which task they want to accomplish.

We are currently in the process of implementing the results of this analysis in an IR system within this domain. We still have a number of challenges ahead of us. First, we need to be able to automatically identify document genre; we have built a training corpus and are refining a method using support vector machines to accomplish this. Second, we need to integrate genre weights into an existing ranking algorithm, and work out a balance between the content and context parameters of the algorithm. Thirdly, we need to determine whether we can successfully collect contextual information with respect to tasks from the searcher in such a way that it will not be overly burdensome.

6. CONCLUSIONS AND FUTURE WORK

This research used a novel method to explore the relationship between work task and document genre in a workplace setting. From this work, we will be able to predict the likely applicability of a given genre once we know the work task and information goal. This, we believe, has the potential to make a significant contribution to the effectiveness of workplace search systems, by incorporating genre weights into the ranking algorithm.

The results from our research are a significant contribution to information retrieval research with respect to the illusive goal of contextual search and the more immediate need for improved enterprise search systems. Identifying and operationalizing contextual variables are significant problems. In this case, we successfully isolated and studied three contextual variables: work task, information task and genre. Although our limited work domain – software engineering consultants – contributed to the success of this work, we believe that a similar relationship among genres and tasks extends to other fields, albeit with different elements of each.

7. ACKNOWLEDGMENTS

This research is supported by an IBM Centre for Advanced Studies Fellowship Grant to the first and second authors and a SSHRC and Canada Research Chairs Program grant to the second author. We would like to thank Julie Waterhouse and Gordon Lee of IBM for their help and support with this project.

8. REFERENCES

[1] T. D. Wilson, "Human information behaviour," *Informing Science*, vol. 3, 2000.

[2] N. J. Belkin, R. N. Oddy, and H. M. Brooks, "ASK for information retrieval: Part I: background and theory," *Journal of Documentation*, vol. 38, pp. 61-71, 1982.

[3] K. Jarvelin and P. Ingwersen, "Information seeking research needs extensions towards tasks and technology," *Information Research*, vol. 10, 2004.

[4] J. e. Allan, "Challenges in information retrieval and language modeling," *SIGIR Forum*, vol. 37, 2003.

[5] L. Freund, E. G. Toms, and J. Waterhouse, "Modeling the information behaviour of software engineers using a work - task framework," presented at American Society of Information Science & Technology Annual Meeting, (accepted), 2005.

[6] A. Broder and A. C. Ciccolo, "Towards the next generation of enterprise search technology," *IBM Systems Journal*, vol. 43, pp. 451-454, 2004.

[7] D. Hawking, "Challenges in enterprise search," presented at Proceedings of the Australasian Database Conference, Dunedin, New Zealand, 2004.

[8] K. Bystrom and P. Hansen, "Work tasks as units for analysis in information seeking and retrieval studies," in *Emerging Frameworks and Methods*, H. Bruce, R. Fidel, P. Ingwersen, and P. Vakkari, Eds. Greenwood Village, CO: Libraries Unlimited, 2002, pp. 239-251.

[9] P. Hansen and K. Jarvelin, "The Information Seeking and Retrieval process at the Swedish Patent- and Registration Office Moving from Lab-based to real life work-task environment," presented at Proceedings of the ACM-SIGIR 2000 Workshop on Patent Retrieval, Athens, Greece, 2000.

[10] J. Yates and W. J. Orlikowski, "Genres of organizational communication: a structurational approach to studying communication and media," *Academy of Management Review*, vol. 17, pp. 299-326, 1992.

[11] W. J. Orlikowski and J. Yates, "Genre repertoire: the structuring of communicative practices in organizations," *Administrative Science Quarterly*, vol. 39, pp. 541-574, 1994.

[12] J. Karlgren, "Non-topical factors in information access," presented at Webnet '99, Honolulu, 1999.

[13] A. Rauber and A. Muller-Kogler, "Integrating automatic genre analysis into digital libraries," 2001.

[14] D. G. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, J. Cai, and X. Liu, "Genre based navigation on the Web," presented at Hawai'i International Conference on Systems Sciences, Maui, Hawai'i, 2001.

[15] N. J. Belkin, G. Muresan, and X.-M. Zhang, "Investigating the effect of the use of user's context on IR performance," presented at Workshop on Information Retrieval in Context (IRiX), SIGIR 2004, Sheffield, England, 2004.

[16] D. He and D. Demner-Fushman, "HARD experiment at Maryland: from need negotiation to automated HARD process," presented at Text Retrieval Conference, Gaithersburg, MD, 2003.

[17] D. W. King, J. Casto, and H. Jones, "Communication by engineers: a literature review of engineers' information needs, seeking processes, and use," Council on Library Resources, Washington, D.C. 1994.

[18] R. S. Taylor, "Question negotiation and information seeking in libraries," *College and Research Libraries*, vol. 29, pp. 178-194, 1968.

[19] M. Hertzum and A. M. Pejtersen, "The information-seeking practices of engineers: searching for documents as well as for people," *Information Processing & Management*, vol. 36, pp. 761-778, 2000.

[20] R. S. Taylor, "Information use environments," *Progress in Communication Sciences*, vol. 10, pp. 217-255, 1991.

[21] R. Fidel and M. Green, "The many faces of accessibility: engineers' perception of information sources," *Information Processing & Management*, vol. 40, pp. 563-581, 2004.

[22] A. Forward and T. C. Lethbridge, "The relevance of software documentation, tools and technologies: a survey," presented at ACM Symposium on Document Engineering, 2002.

[23] M. J. Greenacre, "Theory and Applications of Correspondence Analysis." London: Academic Press, 1984.

[24] H. Askell-Williams and M. J. Lawson, "A correspondence analysis of child-care students' and medical students' knowledge about teaching and learning," *International Education Journal*, vol. 5, pp. 176-204, 2004.

[25] M. Bendixen, "A practical guide to the use of correspondence analysis in marketing research," *Marketing Research On-Line*, vol. 1, pp. 16-38, 1996.

[26] R. Milanese, D. Squire, and T. Pun, "Correspondence analysis and hierarchical indexing for context-based image retrieval," presented at IEEE International Conference on Image Processing, Lausanne, Switzerland, 1996.

[27] S. Nishisato, *Elements of Dual Scaling: an Introduction to Practical Data Analysis*. Hillsdale, NJ: Lawrence Erlbaum, 1994.