JIS

Article

Retrieving haystacks: a data driven information needs model for faceted search

Journal of Information Science
1–18

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Paul Hugh Cleverley
Robert Gordon University

Simon Burnett
Robert Gordon University

Abstract
The research aim was to develop an understanding of information needs characteristics for word co-occurrence based search result filters (facets). No prior research has been identified into what enterprise searchers may find useful for exploratory search and why. Various word co-occurrence techniques were applied to results from sample queries performed on industry membership content. The results were used in an international survey of fifty four practicing petroleum engineers from thirty two organizations. Subject familiarity, job role, personality and query specificity are possible causes for survey response variation. An information needs model is presented: Broad, Rich, Intriguing, Descriptive, General, Expert and Situational (BRIDGES). This may help professionals more effectively meet their information needs and stimulate new needs, improving a systems capability to facilitate serendipity. This research has implications for faceted search in enterprise search and digital library deployments.

Keywords
Enterprise search; digital library; exploratory search; interactive information retrieval; human computer interaction (HCI); information discovery; word co-occurrence; text analytics; oil and gas; information seeking behaviour; serendipity; user interface design.

1. Introduction

Staff in the enterprise have a need to find information which may be located both inside and outside [1] the enterprise. At the same time, the volume of digital information being created in the enterprise is doubling every two years [2, 3]. From surveys, knowledge workers spend approximately a quarter of their time looking for information [4-12] with a significant variance (<10% to >50%) amongst sub-populations [4, 13, 14]. Approximately half of this time spent information seeking may be unsatisfactory in some way [4, 5, 15-19]. There is evidence to suggest staff within enterprises have as much (if not more) difficulty fulfilling their information needs today than they did a decade ago [20, 21]. Data collected from a qualitative survey conducted as part of this research reveals a critical issue in relation to enterprise search:

> Too often retrieve a haystack in which the needles are hard to find. [R_23].

Enterprise search typically refers to technologies which index relevant content (e.g. documents, web pages) to an organization, providing a single place for staff to search without having to know where content resides [22, 23]. Enterprise search is not restricted to unstructured information and can also include structured information within databases. Focusing on the browsing of search results, no prior study has been identified which surfaces what information characteristics of word co-occurrence algorithms enterprise searchers may find useful to browse results and why.

Corresponding author:
Paul Hugh Cleverley, Robert Gordon University, Garthdee Road, Aberdeen, United Kingdom AB10 7QB
p.h.cleverley@rgu.ac.uk

1.1. Framing the research – the bigger picture

Information seeking through human technology interaction may account for less than half of all information seeking activities within an enterprise [24, 25]. This paper is part of a larger body of research examining root causes for sub-optimal information seeking experiences in the enterprise. Some root causes may be related to information searching behaviour, but may include additional aspects of information behaviour such as pre-existing social information practices [26-28] as well as underlying organizational factors such as culture, roles & responsibilities, processes and technological issues. This broader body of work suggests differences between what is documented in corporate policies relating to information searching and use, and the ways in which people actually work with information [29, 30]. Figure 1 shows the scope of this research paper, motivated by the broader need to understand enterprise information needs, with particular focus on exploratory searching where the question is not fully formed in the mind of the enterprise searcher.

The research covers content which may be under the custodianship of a centralized information management service both within (e.g. corporate library) and external to the enterprise (e.g. membership society library or online public access library). Relevant information and literature within the enterprise (but not under central custodianship of a library service) is also within scope. This could include reports and presentations acquired or produced by business teams and placed in an Electronic Document Management System (EDMS) or shared file-system in an informal manner (grey literature).

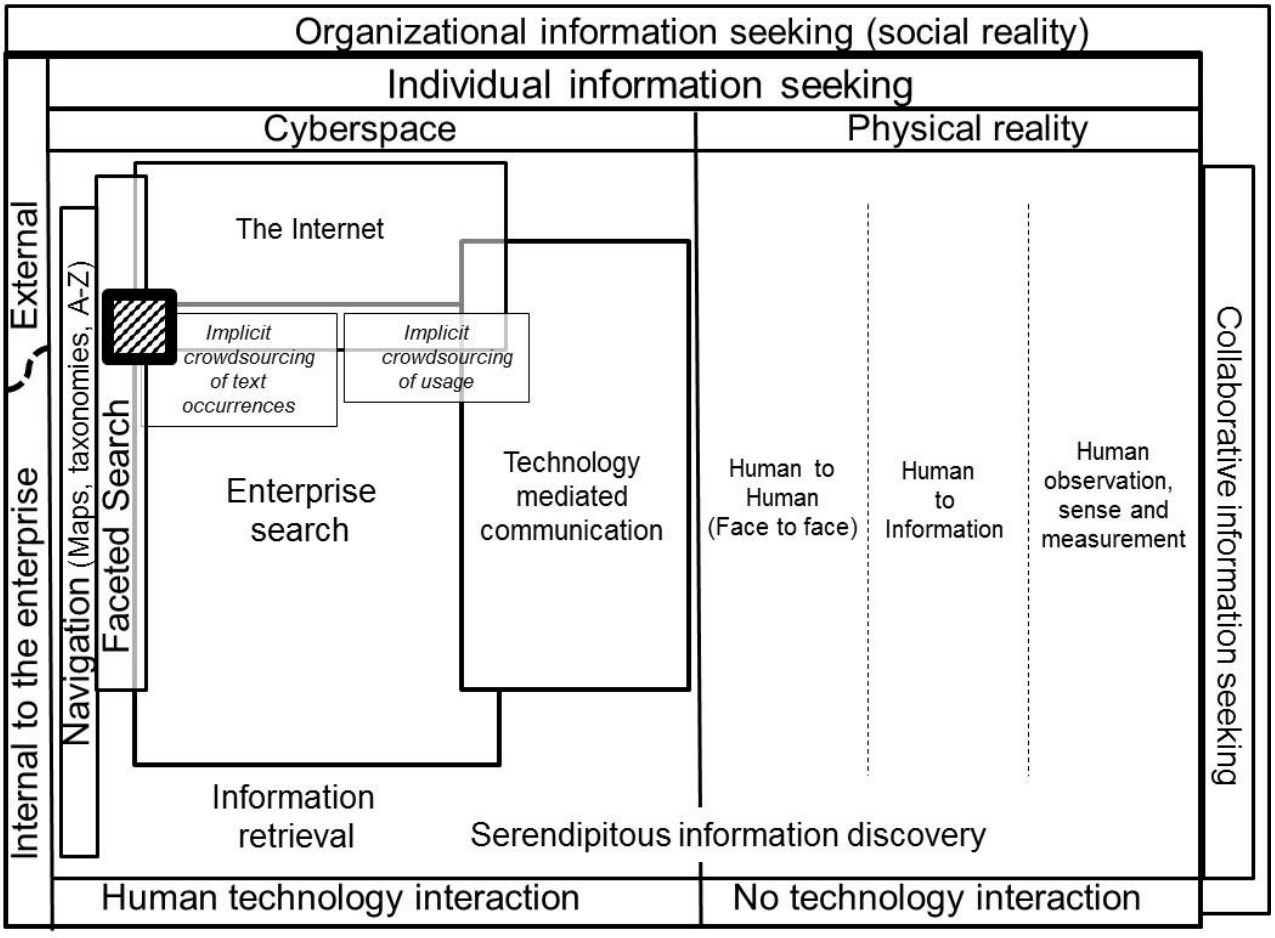

Figure 1. Scope of the research paper (black hashed square) within the context of information seeking in the enterprise.

Crowd sourcing is the practice of explicitly soliciting contributions from a large group of people to aid the completion of a task or provide input to a decision. Implicit crowd sourcing of user activity (collaborative filtering) is

the term given to the use of transactional information to suggest query terms to searchers or recommend information that may be of relevance. In both cases statistical models are created based on end user transactional activity, although each user does not necessarily explicitly know the purposes in which their transactional information will be used.

The crowd within even the largest enterprise is extremely small when compared to the crowd using Internet search engines. Many exploratory technical queries will not have been made previously, so few suggestions can be provided to the user with these techniques. The scope of this paper addresses the implicit crowd sourcing of text (figure 1) within search results to surface associative patterns represented as lists of suggestions (filters) that may aid the browsing of enterprise search and digital library results. This focuses not on end user search activity, but on the words used by authors in proximity to the target search query terms, a concept within the broader field of text analytics.

Content within most enterprises is not web authored with a network of hyperlinks (to the extent content is on the Internet). This may restrict the ability of enterprise searchers to discover information fortuitously. The automatic creation and presentation of parts of such an associative network may mitigate some issues unique to enterprise searching.

Prior research indicates that 70% of all Internet search queries are three words or fewer [31]. With searchers inputting small amounts of query information into increasingly larger corpus sizes, it is perhaps unsurprising that queries may return hundreds if not thousands of results. If the requirement is to locate a 'fact' or single item (i.e. search for a known item) and the relevant information is located on the first page using the classic search box, needs may be fulfilled [32, 33]. However, where a question is not fully formed in the mind of the searcher, such as an exploratory search to learn and investigate (berry picking) [34, 35] additional mechanisms may be required [36-38]. These exploratory searches may account for between 8-20% of all enterprise search use [39]. It is possible the percentage of queries which are exploratory in nature (where the need or question is not fully formed in the mind of the enterprise searcher) may vary between industry sectors. It may also vary if a domain community or search tool (not used by all staff) is analysed in isolation.

1.2. Faceted search

Faceted search is a means of guided navigation or browsing [40], showing a form of grouped categorization (with counts) representing some aspect of what exists in a search result, corpus or website, enabling results filtering (refining). Faceted search allows the same information item(s) to be found through different navigational routes, differing from the folder concept. Faceted search user interface design is the subject of significant and ongoing research interest [41-46] and has shown to be beneficial for exploratory search [47] with some empirical evidence that it can improve user performance [48]. The facet values (typically displayed on the left or top of the user interface) aid the construction of complex high precision queries, which may otherwise be beyond the skill of most users. Facets within user interfaces are:

> The presentation of topic categories on the search user interface to support the refinement of a search query [23]
> Becoming accepted techniques to support complex information seeking tasks like exploratory search [49]

Faceted search is a human computer information retrieval technique [50], an approach which attempts to move searchers from the traditional information retrieval modus operandi: this is what I need towards a greater degree of critical engagement and search intent on the part of the user: is this what I need?

In addition to bibliographic metadata (e.g. author, published date, source), facet types may include names and classifications of 'things' (e.g. document/data types, equipment, techniques, materials). Facet values represent content which has been manually tagged using or automatically classified to (supervised or unsupervised), a controlled vocabulary (master metadata, taxonomy) of terms [51]. Facet values can also represent manual or automatically (supervised or unsupervised) generated metadata not tied to a controlled vocabulary, the former being a folksonomy and the latter a form of data driven entity extraction or clustering.

Some researchers terminologically differentiate between algorithmically generated facet types and values (clustering) and manually created ones (faceted search) [52]. It can be argued that this division is somewhat artificial, as both may produce labelled categories containing results with some form of shared characteristics, so both are in fact capable of enabling the faceted search concept. To support exploratory searching, some researchers advocate clues and signposts are needed for the searcher that are capable of providing the unexpected:

> We cannot dismiss the value of making the search experience enjoyable and even surprising. [53].

Search log data from one digital library indicates faceted search comprises 18% of all search queries [54], although 6% was reported a few years later [55] citing a lack of user awareness. Other studies indicate faceted search usage (interactions) between 5-12% [56, 57]. Focusing on exploratory search queries only, eye tracking studies indicate faceted search may account for 9-17% of what users look at in a search interface rising to 22-29% after a one minute training video [58].

Analysis of the organization providing search log data for this research (150,000 staff conducting 25,000 queries a day) revealed faceted search usage at only 2.7%. Information needs distribution within the organization, information literacy of enterprise searchers, user interface design and uninteresting facets may be potential causal factors for low usage of faceted search.

Faceted metadata typically pertains only to the information item as a whole (e.g. an organizational report), and not to the proximally matched search query contexts within the information item itself. Many clustering techniques [59] apply the same approach to whole documents (or top ranked documents) in search results to produce lists of topics. This is important because where digital information exists in a very narrow technical domain the differences between search results can be quite subtle and contextual. These subtleties can be missed in coarse automatic classification (or clustering) techniques.

Simple word co-occurrence has been shown to aid search query formulation, automatic thesaurus generation [60, 61] and to aid the browsing of search results [62]. While search result overview or facet presentation within the search user interface (e.g. hierarchical facets) or visualization (e.g. spatial word (or collocate) clouds) is not included in the scope of this research [63], it is briefly discussed as some prior studies of co-occurrence facets are tied to word cloud presentations. Word cloud usability has received mixed results in literature. Good usage of co-occurrence word cloud overviews making search faster and aiding discovery have been reported [64], [65], as well as no usage at all [66].

There is some evidence searchers may prefer line-by-line layouts instead of left to right word cloud displays [67]. Searchers may find simple lists of facet terms faster to scan and navigate [68] than spatial word cloud presentations. However, no specific studies test this theory regarding word co-occurrence facets. In general, searchers are treated as a homogenous group, with a one size fits all approach towards faceted search.

2. Method

A mixed methods [69] positivist and constructivist approach was undertaken. This was appropriate given the focus of the research to understand statistically significant preferences and why those preferences were made, without prejudging the possible reasons. An industry survey was used to gather quantitative data on perceptions and corresponding qualitative data on the reasons for choices to gain a detailed understanding of the area. Petroleum engineers in the upstream oil and gas industry were surveyed, using the Society of Petroleum Engineers (SPE) members' online discussion forum so many organizations and geographies could be included.

Those that wished to take part were asked to email the researcher, ensuring participants were motivated. It is acknowledged that voluntary response bias is likely to be present in the results. A semi-structured questionnaire (in Excel) was designed to capture the data. It was piloted on several petroleum engineers before data collection started, and concomitant refinements made. Excel was chosen for this paper based study, as this enabled the participants to complete the questionnaire off-line (without an Internet connection) at their convenience. It is acknowledged that a paper based study would have some limitations, compared with observing how enterprise searchers use facets within an interactive tool within their organizations. Fifty four petroleum engineers (consisting of forty eight men, six women) from thirty two organizations from Europe, North America, Asia and Africa participated.

2.1. Test search queries

To identify aspects of current enterprise information search activity (such as number of words used in a query), analysis of the search logs in a large multinational oil and gas enterprise was undertaken over a three month period. Search query volumes obeyed a power law Zipf distribution, as shown in Figure 2 (squares). The top twenty queries comprise 9.9% of all queries by volume (the head), with a clear body and tail visible, in line with findings from studies in manufacturing industries [70]. The top twenty queries by volume are navigational and/or target information to perform administrative tasks (such as enrolling on a training course, booking a meeting, requesting an Information Technology (IT) service, time-writing and expenses, etc.).

The number of words used in a search query are shown in Figure 2 (triangles) using a rolling ten query average on the right hand (secondary) axis. It was found that 79% of all queries were two terms or fewer. On average, queries made

infrequently have more words than those made frequently. Triangulating with other search log data, the word count spike and lumpiness of the high volume queries may be caused by users choosing the suggested queries in the enterprise search user interface. These implicit crowd sourced queries are suggested using statistical models predicting what the user may wish to search on as they type (often termed auto-complete). This encourages the enterprise searcher to select and make queries with more words than they may otherwise be inclined to.

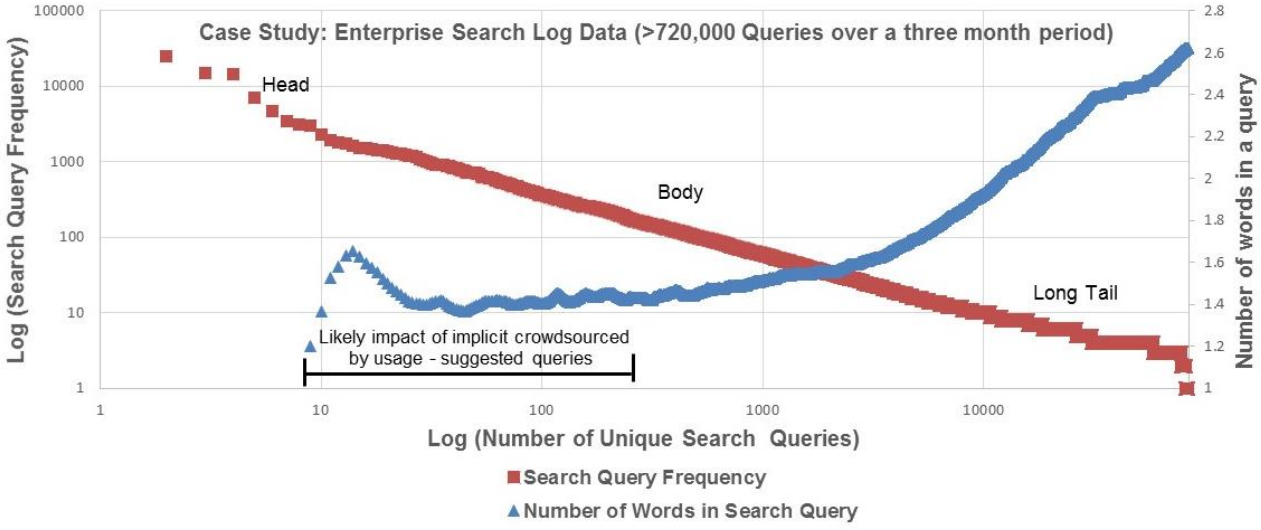

Figure 2. Enterprise search log data (search query frequency and number of words used in a search query) from a large oil and gas enterprise over a three month period.

Technical exploratory queries within the body of the search log in the case study organization were grouped into three representative categories (number of words, taxonomy class and specificity).

An ontology was used for query classification based on a derivative of the Semantic Web for Earth and Environmental Terminology (SWEET) [71] as this suited the subject matter in question. This included the ontologically distinct categories of natural processes/phenomena and materials (matter). These exist in physical reality and it can be argued are independent of whether people are around to witness them. Social categories included human activity (processes/techniques) and mental representations (e.g. common problem). The concept of a 'problem' does not exist in physical reality, only in social reality. It is therefore subject to the complexities associated with the social construction of knowledge [72] and like all knowledge can change through time. What is reality for one individual or group is not necessarily the same for another, leading to the formation of sub-cultures with shared beliefs. These individuals and groups use digital library and enterprise search systems, so it follows that the end users of these system may not have homogeneous needs.

A category 'common problem' was created for the study. For simplicity, search query terms were only classified to this category where they were clearly and universally acknowledged as common problems within the petroleum industry. The PetroWiki of SPE, based on the Petroleum Engineering Handbook (PEH) [73], was used as a basis to aid this classification. The SPE is the largest individual member organization (over 120,000 members) serving managers, engineers, scientists and other professionals worldwide in the upstream segment of the oil and gas industry. For example, 'stuck-pipe' and 'corrosion' query terms from the search log were classified as a 'common problem' as they are widely acknowledged as common problems in the industry. For 'stuck pipe' and 'corrosion' respectively:

> During drilling operations, a pipe is considered stuck if it cannot be freed from the hole without damaging the pipe, and without exceeding the drilling rig's maximum allowed hook load. Pipe sticking can be classified under two categories: differential pressure pipe sticking and mechanical pipe sticking. Complications related to stuck pipe can account for nearly half of total well cost, making stuck pipe one of the most expensive problems that can occur during a drilling operation. [73]

> Corrosion of metal in the presence of water is a common problem across many industries. The fact that most oil and gas production includes co-produced water makes corrosion a pervasive issue across the industry. Age and presence of corrosive materials such as carbon dioxide (CO2) and hydrogen sulphide (H2S) exacerbate the problem. [73]

The query 'corrosion' is also a natural process/phenomena (as well as causing a common problem). Another classification type for search queries was based on specificity. In this context, term specificity is defined as its frequency of occurrence within the corpus (the more frequent a word occurs the less specific it is). Specificity was determined by counting the number of occurrences of the term in the SPE dataset, this was triangulated with results from the Google n-gram viewer [74] that counts word frequency in millions of books.

Initial testing with several enterprise staff, indicated because practicing business professionals were being targeted with limited time at their disposal, the survey should take no more than fifteen minutes to complete otherwise staff would not spend the time completing the survey. In traditional studies of information retrieval [75] it is common to use fifty search queries (or more) to compare different algorithms or technologies to minimize the comparison error. The literature [76] supports the use of a smaller number (e.g. four search terms) to elicit responses to test certain situations.

It was felt that four different search terms was sufficient to elicit responses related to search term taxonomic class and term specificity.

It is acknowledged that only testing four queries is a limitation of the research. The four (exploratory in nature) queries chosen were, 'corrosion', 'reservoir management', 'stuck-pipe' and 'shale gas'. These were purposefully sampled from different taxonomic groups within the enterprise search log ensuring taxonomic class and term specificity diversity.

2.2. Test corpus and co-occurrence algorithms

Agreement was given by the SPE to use their 70,000 abstracts for this research, generating a corpus of over one million words. This simulated a sub-set of an enterprise corpus relevant to the field of enquiry.

The distributional hypothesis [77] states words appearing close to each other in text, are likely to be associated (share meaning) in some way. Text collocations are statistically significant text co-occurrences in a direct syntactic relation [78]. A key parameter is window size around target query terms. A window span size of two to ten terms is often proposed as having cognitive plausibility [79, 80], although larger spans of 50-100 have also been proposed [81, 82]. A window size of sixteen words was chosen as only abstracts were available. Python [83] scripts were written to extract the text associations, common stop words (e.g. then, and) were filtered out [84].

In order to elicit a wide range of responses regarding enterprise searcher needs, a diverse set of word co-occurrence term characteristics was required for the four search queries. In this context, 'characteristic diversity' was achieved through single and multi-word terms, differing word specificity (frequency of occurrence) and semantic relatedness diversity (not just semantically similar 'is-a' or 'part-of' terms). It is acknowledged that other measures could increase diversity characteristics (e.g. word length/number of characters) but were not included in this study. During the iterative process of grounded theory, a further diversity element was introduced based on adjacency versus non-adjacency for multi-word terms (see Section 3.2).

After a brief review of the literature, three first order co-occurrence n-grams were chosen to generate the navigational suggestions, as they delivered the desired range of diversity characteristics described above. Other algorithms may also deliver similar diversity characteristics for co-occurring terms but it was not in the research scope to compare algorithms. The three n-grams were as follows:

- A unigram ranked by descending frequency (List A) for the single words most commonly associated to the query term(s) within the text;
- A bigram ranked by descending frequency (List B) that listed the word pairs most commonly associated to the query term(s). For the bigram, an extra post processing algorithm was applied, removing all bigrams that contained any mention of the query terms. For example, for the query 'corrosion', the bigram 'metal corrosion' was removed, avoiding 'taxonomic type variations' addressed by existing autocomplete techniques;
- List C is as List A, except ranking is by pointwise mutual information measure [85]. This favoured discriminant co-occurring terms often containing non-obvious associations.

The algorithms were used to elicit user information needs characteristics as they produced sufficiently diverse results. It is not suggested that these are necessarily the best co-occurrence algorithms and it was not in the research scope to compare algorithms, the focus was on testing the enterprise searcher response to different characteristics.

A review of the enterprise search user interface used within the case study organization, along with a selection of other online search tools (including the SPE digital library) provided some evidence that facet lists are often truncated perhaps arbitrarily after the first twenty terms (which are often ordered by frequency within each facet). In some cases only the first five terms are visibly shown by default within the faceted search user interface probably due to screen space limitations.

To test the extent values lower down in the ranking order (not normally seen by enterprise searchers) may be useful, thirty co-occurring terms were used in the survey. The results for the 'stuck-pipe' query flat lists (hierarchies not tested) of the top thirty terms given to respondents in Microsoft Excel are shown in Table 1. Conceptually, the lists are vectors on a directed graph network, choosing a value would narrow (filter) existing search results and display new text co-occurrences, traversing the network.

Table 1. Ranked navigational filter suggestions for the 'stuck-pipe' query for the three different n-gram techniques.

Rank	List A	List B	List C
1	Drilling	lost circulation	differentially
2	Problems	problems such	freeing
3	Hole	well control	spotting
4	Lost	poor hole	incidents
5	Incidents	hole instability	sticking
6	Well	hole cleaning	risked
7	Risk	drilling operations	troubles
8	Cost	freeing differentially	jarring
9	Loss	while drilling	caving
10	Circulation	tight hole	sloughing
…	…	…	…
30	Reduced	open hole	costly

The thirty values for each of the four queries were displayed on a separate tab in Excel. The thirty values were displayed in blocks of ten, separated by a blank line so they were easier to read.

2.3. Respondent ranking of algorithms

The survey respondents were asked to rank the navigational suggestion lists in order of usefulness for each given query and explain why this order was felt to be appropriate. The task was based on exploratory searches, to learn and investigate about the topics represented by the four search query terms. As the task was exploratory in nature (the question was not fully defined, as an exploratory search would not be fully defined in an enterprise searchers mind) it would be possible for participants to arrive at different intents (within the task scope) based on their own personal context and the stimulus provided by the word associations. As the respondents targeted in the survey belonged to the same industry and subject discipline, it was anticipated that many intents would probably be the same across the sample. Response variance was of interest and the reasons given by the respondent for their choices.

To avoid central tendency bias and include expressiveness, a collective relative ranking of the generated lists was chosen per query, covering an empty set (Ø), equivalence (equal rank) and partially ordered sets (poset). Respondents could therefore omit distracting or unhelpful navigational suggestion lists. Using factorial equations, the possible permutations (P) for {A,B,C} allowing for an empty set, equivalence and partial sets is twenty six.

2.4. Data analysis

The comments were analysed and coded simultaneously in real time, as each completed survey was received, using an immersive approach based on grounded theory [86] and axial coding [87], moving from description to conceptual models, keeping an open mind but not an empty head [88]. As themes emerged, further information was sought from the survey respondents. Contextual data were collected from respondents (e.g. experience, location, company type).

An issue visualizing ordered data is how to avoid introducing visually misleading trends on a standard histogram. One solution (which was used in this research) is to use a Permutohedra visualisation [89]. This treats the rankings as points in n-dimensional Euclidian space, the Permutohedra appearing as the convex hull of the rankings.

3. Results

A coded system was used to denote survey respondents for anonymity, for example [R_1] is respondent number one. All respondents found at least one suggestion list useful, in 56% of cases all three lists were found to be useful. Survey respondents indicated that combined, the lists hint at the 'big picture' of the search topic.

3.1. List A

List A was characterized by its broad theme and general nature by respondents, illustrated by the following representative comments which combine both positive and negative opinions:

Topic container,
Easy to understand,
Caters to wide audience [R_10, R_27, R_8]
Too general [R_9, R_25, R_31, R_44]
List A seems to get straight to the point [R_19]
A was far more general although some of the terms towards the end of the list…would be of interest. [R_32]
Quite dry [R_4].

Experts working across industry sectors found List A useful to disambiguate subject areas. Some subject matter experts found the ten most frequent suggestions in Lists A and B 'relevant but not interesting', finding some of the richer terms lower down the frequency ranked lists of greater interest. Where a preference was expressed, as many terms of interest fell outside the top ten, as fell within it. This supports findings from a previous study where users found the most frequently occurring (or most popular) words to be of less interest [63]. This is significant as many enterprise search and digital library interfaces tend as a default to show the top five or ten terms within a facet ranked by frequency of occurrence. By doing so they may overlook a latent information need for the more unusual (less frequent) categories present in their search results. A theme based on richness and diversity of terms was identified.

3.2. List B

List B was the most popular, ranked first in 75% of all cases, and characterized by the following comments:

Descriptive [R_15, R_25, R_42]
Meaningful [R_10]
Instructive [R_3]
Right context [R_21]
Specific enough for my level [R_31]
Capturing range of contexts with two word summaries [R_10]
Two words better than one [R_10, R_21]
List B's multi word approach won hands down [R_14].

As this multi-word descriptive theme emerged, as part of the iterative process of investigation, it was explored further with ten of the respondents. More complex forms of word co-occurrence were used (termed 'Topic modelling') [90-92]. This probabilistic statistical technique is capable of surfacing latent associations between words which are not explicitly adjacent (contiguous) to each other in the original text. This makes the lists it produces fundamentally different from the n-gram model (e.g. bigram, trigram) which is based on the adjacent nature of words in a set window. The Topic modelling techniques were used to generate two, three and four word lists (D, E and F respectively) which can be seen in Table 2. Respondent comments for D, E and F included the positive and negative themes illustrated by:

Surfacing a scenario I had not thought of [R_14]
Incomprehensible and confusing [R_45, R_53].
Algorithms D, E and F contain some "interesting" words, but combinations are pretty strange/random-looking. So, while there are more "interesting" words cropping up it's very difficult to know how the assessment of "interestingness" is influenced since you have a conflict between having more "interesting words" but, on the negative side, their juxtaposition is strange so detracts. [R_36].

Comparing the multi-word theme, respondents preferred List B (to D, E or F) as it was deemed more coherent for filtering their search results. However, there may be potential where the enterprise searcher is open to spending more time interpreting the multi-word associations to make more insightful inferences about the underlying topics. If the enterprise searcher does not have the time to spend on interpretation, it appears they are more likely to consider it distracting.

For the search query 'reservoir management' (a human activity (process, technique)) List B was ranked first almost exclusively by the entire population (Figure 3). This homogeneity contrasts sharply with the results from the other three queries.

The other queries that consisted of two words displayed more response variation e.g. 'shale gas', so the number of words in the query is unlikely to be the only causal factor. Term specificity (how frequent the word is mentioned in literature) is also unlikely to be a cause, as 'stuck-pipe' is more specific and 'corrosion' is a far less specific term (relative to 'reservoir management') and both display more response variation. It is possible the taxonomic class of the query could play a role, with terms related to human activity eliciting a different response to those that refer to other taxonomic classes (e.g. natural process/phenomena, materials (matter)). However there is little evidence that taxonomic class plays a role based on the two queries with a shared taxonomic class in this study ('stuck-pipe' and 'corrosion' search queries sharing the same classification of 'common problem').

There may be a relationship to the number of words used in concept names in each taxonomic space, however there was insufficient data in this study to test this hypothesis.

Table 2. Navigational suggestion lists (D, E and F) for the 'stuck pipe' query using Topic modelling techniques with two, three and four words respectively.

No.	List D	List E	List F
1	pressure pore	salt problems due	pressure formation control pressures
2	tubing nbsp	drilling casing wells	pipe string point bottom
3	sticking differential	system control operation	time operations process real
4	drilling system	stuck-pipe cost data	drill-string torque hole drag
5	drilling gas	barite sag shear	drilling fluid flow cleaning
6	drill string	oil gas wells	drilled formation depth surface
7	wells reservoir	stuck operations point	drilling circulation lost mud
8	impact string	liner depleted set	cuttings transport wellbore cutting
9	pipe liner	flow parameters model	drilling system high performance
10	drilling casing	mud cake drill	oil based water muds
…	…	…	…
30	fluid results	hole drilling offshore	offshore rig cost project

The thirty values for each of the four queries were displayed on a separate tab in Excel. The thirty values were displayed in blocks of ten, separated by a blank line so they were easier to read.

Another explanation may relate to the linguistic diversity of the terms natural taxonomic hierarchy. The term 'reservoir management' is a broad umbrella term for a whole host of business processes and techniques, relating to data acquisition, subsurface and economic description, modelling and simulation, production forecasting and reservoir surveillance. Few if any of the concepts (e.g. 'production forecasting' and 'reservoir surveillance') subsumed by the term 'reservoir management', explicitly contain the phrase 'reservoir management' as part of their phrase description. Therefore occurrences of these techniques in the text would have been missed by the algorithmic techniques which applied a relatively crude keyword matching approach.

Despite 'corrosion' being a less specific term, it contained far more specific terms within its word co-occurrence lists than reservoir management. One causal factor may be due to the concepts subsumed by the term 'corrosion' contain the term 'corrosion' within their phrase description (e.g. 'fretting corrosion', 'high temperature corrosion' and 'flow-assisted corrosion').

In summary, it is possible the algorithms did not capture a true representation of the concept of 'reservoir management' from the SPE corpus. This may have combined with other factors such as the number of words used for

concept descriptions (within the associated subsumed concept hierarchy) downplaying the usefulness of the single word suggestions within Lists A and C.

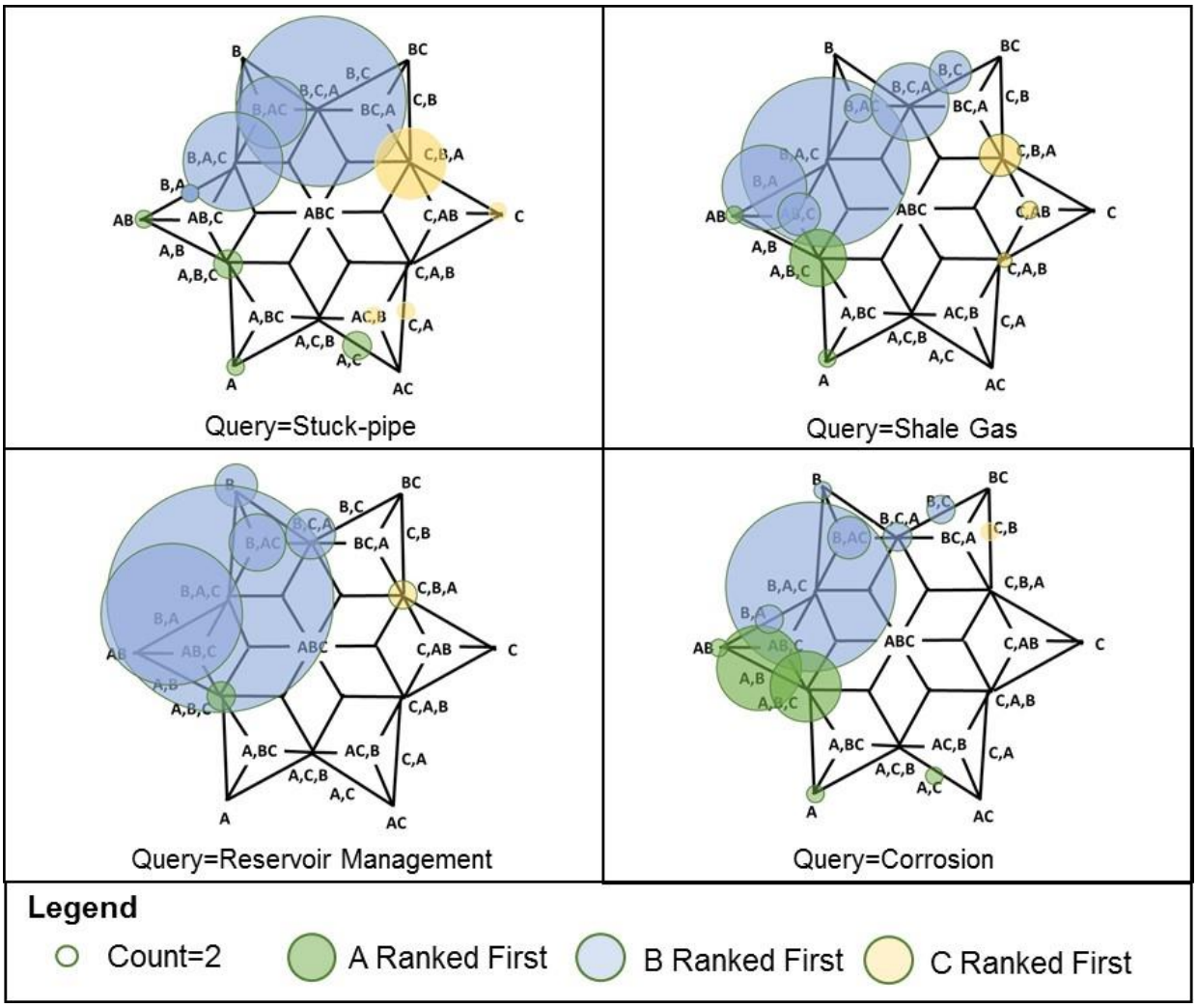

Figure 3. Respondent preference permutations for navigational suggestion lists represented on a Permutohedra visualization. Bubble size is related to number of respondent votes, Where List A is ranked first bubbles are coded green, where List B is ranked first bubbles are coded blue and where List C is ranked first bubbles are coded orange. Sample size=54.

3.3. List C

List C was generally seen as too specific by most respondents, but elicited some interesting variations and observations related to new knowledge acquisition and serendipity. A theme based on intrigue emerged. Clearly what one person may find intriguing, another person may not. However, in general it appears some algorithmic techniques may be more likely than others to surface terms (in proximity to their query terms within their search results) that people may find intriguing:

Too specific [R_44, R_53, R_9]
Obscure [R_39]

Too intimate for general engineer [R_54]
Useful for detailed dives [R_31]
New vocabulary might learn something [R_32]
Purely intriguing high on "interestingness" quotient, you can't say where these search results could lead you [R_8].
I guess it's a trade-off between novice and advanced users [R_14]
Answers I would suggest will be dominated by the level of the reviewer. If I am a detailed subject matter expert I would answer C first [R_31].

3.4. Contextual differences

Two different preference definitions were used for statistical inference of the categorical data; lists ranked first and lists ranked first or second (Table 3), the latter enabling the surfacing of subtle patterns masked by the dominance of list B. Subject matter experts showed preferences for list C supported by survey comments, identifying a theme based on expertise. However this result should be treated with caution due to the marginal level of statistical significance.

No relationship was found between respondents from an oil and gas operator compared to an oil and gas service organization. Respondents with explicit research roles (different disciplines) displayed preferences for list C as first or second choice (Table 3) which is statistically significant. This may indicate that job role is an influencing factor on algorithmic choices. There was insufficient clean data to test national culture, which was further complicated by the presence of expatriate staff in the survey. Previous studies do indicate a link between culture and information seeking behaviour [93, 94].

Table 3. Chi square test for independence p-values within sub-populations.

Sub-population	Ranked 1st	Ranked 1st or 2nd
Significance test: Total population	<0.001	<0.001
Role: Operator company group	0.99	0.83
Role: Service company group	0.75	0.83
Personality: Top 20% word count group	ef <5	0.03
Personality: Bottom 20% word count group	ef <5	0.01
Specificity: Stuck-pipe query responses	ef <5	<0.001
Specificity: Shale gas query responses	0.86	0.87
Specificity: Reservoir management query responses	<0.001	0.07
Specificity: Corrosion query responses	<0.001	0.05
Subject familiarity: stuck-pipe (A and B) v C*	0.05	0.28
Role: Research company group (A and B) v C*	0.60	0.01

Chi square parameters used where (A or B or C) was chosen: significance level (alpha)=0.05, degrees of freedom=2, *N*=205.
* Fisher test used (alpha=0.05) where some expected frequencies (ef) <5 when testing 1st and (1st or 2nd) ranked categories

There was a significant degree of variation in word count within respondent survey comments. At the low end of the count, the bottom 20% gave an average of three word answers, at the high end, the top 20% gave an average of eighty two word answers. According to [95] the amount of information provided by a respondent may depend on the respondent's personality. The top 20% showed preferences in all four queries for list C, the bottom 20% preferences for list A (Table 3). An example quote from the top group:

I like C very much as it tackles some of the more "soft" issues that regularly occur in actual business (outside world of theory), such as conflict and workflows. A, I did not like, too vague and no promise of telling me anything I didn't already know' [R_37].

It may be possible to infer personality influences information searching activity [96], [97]. This could present further research opportunities on relationships between personality (e.g. openness to experience-intellectual curiosity), faceted search algorithms [98] and serendipity, highlighting intriguing associations [99].

3.5. Query specificity

Approximately half (52%) of all cases where C was ranked first, were for the narrowest (most specific) search query 'stuck-pipe' and almost half (48%) of all cases where A was ranked first were for the broadest (least specific) query 'corrosion' (Figure 3, Table 3). The data could imply that certain algorithmic methods for co-occurrence suggestions are more appropriate than others, based on the specificity of the query terms used. For word frequency, [100] indicate counts lower than the seed query may be easy to interpret. However, where 'very large' subsumption gaps exist between the query term and co-occurring word frequency (Figure 4), single word terms may be hard to interpret for those unfamiliar with a subject and judged over specific (with a greater risk of being deemed not relevant to the initial information need) by subject matter experts.

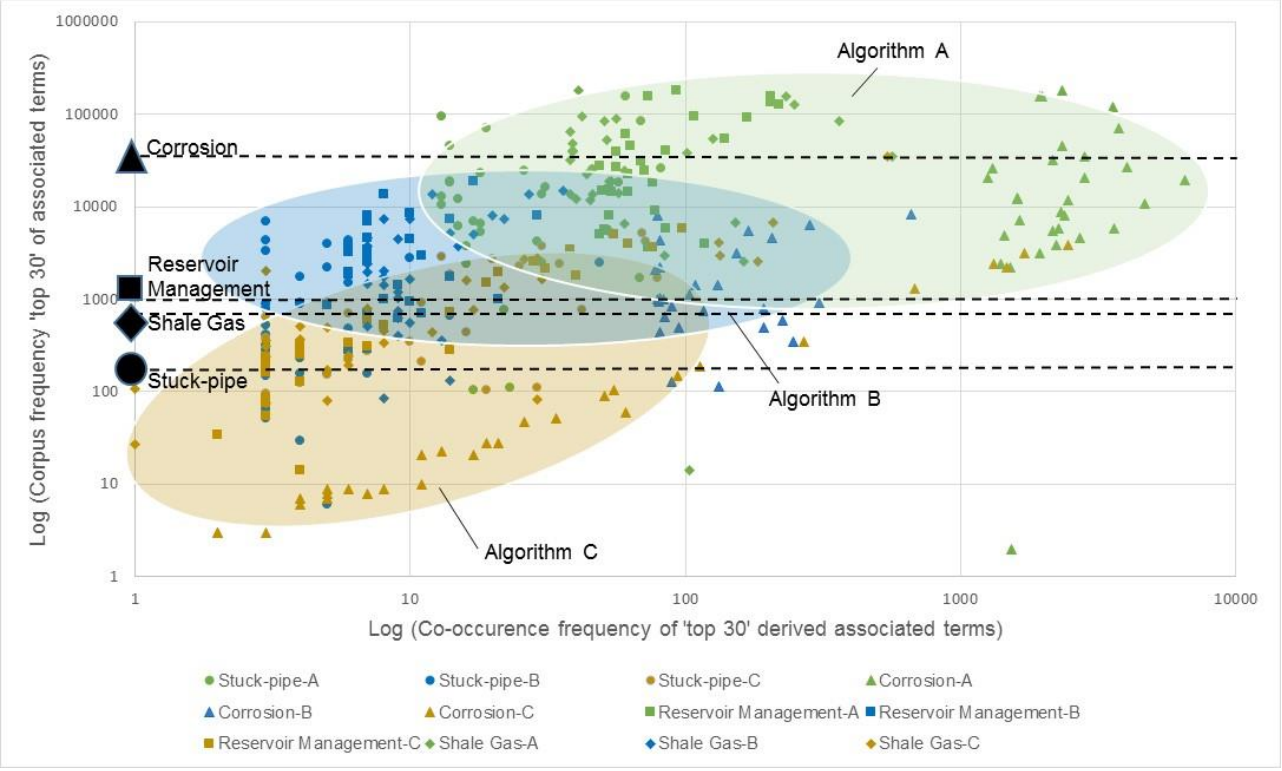

Figure 4. The relationship between the frequency of co-occurring terms (x-axis) and their frequency in the dataset as a whole (y-axis) for each of the four search queries (which are also plotted on the y-axis by their own frequency in the dataset) for the three navigational suggestion Lists A, B and C.

3.6. Situational need

An information needs based on temporal situational examples (instances) emerged. Comments for this situational theme include the following, which were made for Lists A, B and C:

> I would be looking for incidents [R_16]
> List A grabbed my attention with "incidents, events, occurrence", because I want to know more about what exactly went wrong [R_38]
> List B could lead to more knowledge acquisition.. terms like "case study" could lead you to a place ..learn so much more [R_8].

One respondent mentioned temporal (information currency) effects as a reason why they chose List C, although temporal effects were not explicitly tested during the study. In a follow up interview with [R_53], the dislike of repeated

terms was raised (e.g. hole instability, wellbore instability) where algorithms did not group pseudo-synonyms. The use of second order co-occurrence techniques could possibly help to eliminate some of these issues.

3.7. Development of an information needs model

The themes which have emerged from the respondent's comments identified in the previous sections (3.1-3.6) can be combined to form a model to identify information characteristics which may be advantageous for word co-occurrence search result filters. A review of existing literature [46, 97, 101-108] in combination with findings derived from the primary data in this research has aided the development of a model for word co-occurrence in faceted search (Table 4). This consists of the seven user needs; Broad, Rich, Intriguing, Descriptive, General, Expert and Situational (BRIDGES). This model could be used to complement existing controlled vocabularies and other facets and mechanisms such as social signs, currency and provenance metadata. The BRIDGES model will be explored in more detail in a future article.

Table 4. BRIDGES word co-occurrence information characteristics needs model for faceted search filters.

Facet Need	Description
Broad	Road signpost analogy. Large container topics, helicopter overview for navigation. Help disambiguate between industry sectors and those unfamiliar with a subject.
Rich	Relevant, comprehensive and diversified set of suggestions not just the most frequent/popular. Concrete, abstract, divergent, emotive (sentiment) terms, synonyms/acronyms.
Intriguing	Interesting, engaging, divergent, unusual, non-obvious terms (or term combinations), which may lead to unanticipated or surprising results.
Descriptive	Multi-word theme which is meaningful, expressive, logical words which describe, instruct and inform the searcher. Clear not ambiguous. Coherent not distracting or disjointed.
General	General store analogy. Right level for searcher, accurate terms they can relate to close to search terms specificity. Everyday parlance language.
Expert	Boutique store analogy. Focused, specialist, narrow, theory, specific terminology. Recognizable by subject matter experts, not generalists.
Situational	Real world informational examples in space and time. Events, instances, incidents, case studies. Named entities in context: e.g. products, people, places, projects, organizational.

The Broad, General and Expert parts of the model relate to the information need based on the level of the enterprise searcher in terms of subject experience and the need to disambiguate between broad subject topics. The descriptiveness theme relates to the semantic clarity of the word co-occurrence facet. The richness and intriguing parts of the model relate to the diversity, and volume of terms which could increase the propensity of word co-occurrence facets to facilitate serendipitous information encounters. Finally, the situational theme surfaces instances of activity or names of things present in the search results, that cuts across all of the themes mentioned above.

4. Conclusion

The research has shown that enterprise searchers in an oil and gas context may find word co-occurrence suggestions useful as a potential navigational filter aid (set of prompts) to browse search results. There were clear preferences for certain word co-occurrence characteristics based on a number of contextual factors. Contextual differences such as subject familiarity, job role, personality and search query specificity are probable causal factors for response heterogeneity.

A review of existing literature [23, 46, 48, 65-67] and mainstream search engine user interfaces [109-112] appears to indicate some deployments are not meeting all the information needs of the BRIDGES model within their facets. These information systems are therefore potentially missing some of the exploratory information needs of its users. Whether the findings from this study in the oil and gas industry are applicable to other industry sectors is uncertain.

An understanding of the BRIDGES information needs model may lead to more effective fulfilment of the underlying need that drove the user to make a search and/or act as a catalyst to stimulate (through intriguing associations) a related, but entirely new information need, improving a systems capability to facilitate serendipity. In particular, presenting intriguing terms within facets of search interfaces to surprise the user, could lead to leaps in value which are impossible

to predict. These connections may far outweigh the cost reduction based Return on Investment (ROI) approaches traditionally employed to justify investments in search tools [113]. A key factor is how word co-occurrence filters are implemented, to minimize distraction (they may not be useful for all search modes), whilst offering potential serendipitous opportunities. This research has implications for digital library and enterprise search engine deployments. It is possible with improved information literacy, searchers may utilize co-occurrence facets more effectively if they understand how they work and the value they could potentially bring to their exploratory searches.

Although end user interactions with variants of List A have been studied before to some extent [47, 48] it is believed that interactions with List B and C have not. The research is limited as it simulated facet values by using lists in Microsoft Excel. It is possible if respondents were able to click through to actual results through a search user interface, some responses may differ. Potential areas of further research could include; testing these findings in an interactive search user interface, use of trigrams (insufficient data in this study but may be of value as filters [104]), different window sizes, facet ranking techniques and use of colour to present terms by various categories.

Acknowledgements

Appreciation is given to the Society of Petroleum Engineers (SPE) and the oil and gas organization (to remain anonymous) which provided case study data and all the petroleum engineers and scientists who participated in the survey.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

[1] De Saulles M. Information literacy amongst UK SMEs'. An information policy gap. Association for Information Management (ASLIB) proceedings: New Information Perspectives 2006; 59(1): pp. 69-79.

[2] Gantz J and Reinsell D. Extracting Value from Chaos: Report ID 1142. http://www.emc.com/digital_universe (2011, accessed January 2014).

[3] Gartner. Information Management Goes 'Extreme': The Biggest Challenges for 21st Century CIOs. http://www.sas.com/offices/NA/canada/lp/Big-Data/Extreme-Information-Management.pdf (2011, accessed March 2014).

[4] Doane M. Cost benefit analysis: Integrating an enterprise taxonomy into a SharePoint environment. Journal of Digital Asset Management 2010; 6(5): pp. 262-278.

[5] International Data Corporation (IDC). The high cost of not finding information. http://www.ejitime.com/materials/IDC%20on%20The%20High%20Cost%20Of%20Not%20Finding%20Information.pdf (2001, accessed March 2013).

[6] Adkins S. Information Gathering in the Electronic Age: The Hidden Cost of the Hunt. Safari Techbooks Online (2003, accessed November 2013).

[7] Search Technologies. Online Article http://www.searchtechnologies.com/enterprise-search-surveys.html (accessed December 2013).

[8] Webster M. Bridging the information worker productivity gap: New challenges and opportunities for IT. http://www.adobe.com/uk/products/acrobat/idc-bridging-productivity-gap-white-paper.html (2012, accessed January 2014).

[9] Lowe A, McMahon C, Culley S. Characterising the requirements of engineering information systems. International Journal of Information Management 2004; 24: pp. 401-422.

[10] Delphi Group. Taxonomy and Content Classification. Market Milestone Report. www.delphigroup.com (2002, accessed January 2014).

[11] Wipro and Oil & Gas Journal. Oil and Gas Upstream Data and Information Management Survey 2011. http://www.wipro.com/Documents/insights/oil_and_gas.pdf (accessed November 2013).

[12] McKinsey and International Data Corporation (IDC). Impact of Internet Technologies: Search 2012. http://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/High%20Tech/PDFs/Impact_of_Internet_technologies_search_final2.aspx (accessed March 2013).

[13] Majid S, Anwar MA and Eisenshitz TS. Information need and information seeking behaviour of agricultural scientists in Malaysia. Library and Information Science Research 2000; 22(2): pp. 145-163.

[14] Findwise. 2013 Findability Survey (Stockholm May 2013). http://www2.findwise.com/findabilitysurvey2013 (accessed December 2013).

[15] Geosoft. Exploration Information Management survey 2011. http://www.geosoft.com/media/uploads/resources/brochures/exploration_information_management_survey.pdf (accessed January 2014).

[16] International Data Corporation (IDC). The hidden costs of information work, 2005. http://www.slideshare.net/PingElizabeth/the-hidden-costs-of-information-work-2005-idc-report (accessed April 2013).

[17] International Data Corporation (IDC). Information advantage: Information access in tommorow's enterprise, 2009. http://www.slideshare.net/PingElizabeth/the-information-advantage-information-access-in-tomorrows-enterprise (accessed November 2013).
[18] Association for Information and Image Management (AIIM). Findability: The art and science of making content easy to find, 2008. http://www.aiim.org/Research-and-Publications/Research/Industry-Watch/Findability (accessed November 2013).
[19] Microsoft and Accenture. Upstream oil and gas computing trends survey (conducted by PennEnergy and the Oil and Gas Journal Research Centre), 2010. http://www.accenture.com/us-en/Pages/insight-upstream-oil-gas-trend-survey-2010-summary.aspx (accessed November 2013)
[20] Stratigos A. Online resource. http://www.businesswire.com/news/home/20050510005823/en/Outsell-Survey-Shows-Knowledge-Workers-Turning-Open#.U4BNkdFeHIU (2005, accessed January 2014).
[21] Mindmetre. Mind the enterprise Search Gap, 2011. http://www.smartlogic.com/home/knowledge-zone/white-papers/1600-mindmetre-research-report-sponsored-by-smartlogic (accessed March 2013).
[22] Hawking D. Challenges in Enterprise Search. Fifteenth Australasian Database Conference (ADC) 2004, Dunedin, NZ. Conferences in Research and Practice in Information Technology 2004; 27: pp. 15-26.
[23] White M. Enterprise search. 1st ed. United States: O'Reilly, 2012: pp. 164.
[24] Robinson MA. An empirical analysis of engineer's information behaviors. Journal of the American Society for Information Science and Technology 2010; 61(4): pp. 640-658.
[25] Burnett S. The strategic role of knowledge auditing and mapping: An Organizational Case Study. International Forum on Knowledge Asset Dynamics (IFKAD)-Knowledge Cities World Summit (KCWS) 2012, 12-15th June Matera, Italy.
[26] Wilson, TD. Models in information behaviour research. Journal of Documentation 1999; 55(3): pp. 249-270.
[27] Kling R and McKim G. Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science 2000; 51(14): pp. 1306-1320.
[28] Bates MJ. Information Behavior. In: Bates MJ, Maack MN (eds). Encyclopedia of Library and Information Sciences. 3rd ed. New York: CRC Press, 2010, pp. 2381-2391.
[29] Brown JS and Duguid P. Organizational learning and Communities of Practice: Towards a unified view of working, learning and innovation. Organizational Science 1991; (2)1.
[30] Stamper R. Signs, Information Norms and Systems. In: Holmqvist (ed). Signs of Work, Semiosis and Information Processing in Organizations. 1st ed New York: Walter de Gruyter, 1996.
[31] Taghavi M, Patel A, Schmidt, N. An Analysis of web proxy logs with query distribution pattern approach for search engines. Computer Standards and Interfaces 2011; 34 (1): pp. 162-170.
[32] Chilton LB and Teevan J. Addressing People's Information Needs Directly in A Web Search Result Page. World Wide Web (WWW) 2011 Proceedings of the 20th International Conference on World Wide Web 2011: pp. 27-36.
[33] Russell-Rose T, Lamantia J and Burrell M. A Taxonomy of Enterprise Search. Human Computer Information Retrieval (HCIR) 2011; 763 (Session 1): pp. 15-18.
[34] Bates MJ. The design of browsing and berrypicking techniques for the online search interface. Online information review 1989; 13 (5): pp. 407-424.
[35] Pirolli P and Card S. Information foraging in information access environments. Computer Human Interaction (CHI) 1995 Proceedings of the SIGCH Conference on Human Factors in Computing Systems 1995: pp. 51-58.
[36] Aula A and Russell DM. Information Seeking Support Systems. A workshop sponsored by the National Science Foundation (NSF), Capitol Hill June 26-27 2008. Complex and Exploratory Web Search. http://ils.unc.edu/ISSS/ (2008, accessed January 2014).
[37] Morville P and Callendar J. Search Patterns. 1st ed. Canada: O'Reilly, 2010.
[38] Ellis D and Haugan M. Modelling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation 2009; 53 (4): pp. 384-403.
[39] Stenmark D. Identifying clusters of user behaviour in Intranet Search Engine Log Files. Journal of the American Society for Information Science and Technology 2008; 59 (14): pp. 2232-2243.
[40] Ellis D. A Behavioural approach to Information Retrieval and System Design. Journal of Documentation 1989; 45(3): pp. 171-212.
[41] Hearst M. Design Recommendations for Hierarchical Faceted Search Interfaces. In the Association for Computing Machinery (ACM) Special Interest Group for Information Retrieval (SIGIR) Workshop on Faceted Search 2006.
[42] Wilson ML, Andre P and Schraefel MC. Backward Highlighting: Enhancing Faceted Search. Proceedings of the 21st Symposium on User Interface Software and Technology (UIST2008), October 19-22, 2008, Monterey, CA, USA: pp. 235-238.
[43] Clarkson EC, Navathe SB and Foley JD. Generalized formal models for faceted search interfaces. Joint Conference on Digital Libraries (JCDL) 2009 Proceedings of the 9th Association for Computing Machinery (ACM)/ Institute of Electricals and Electronics Engineers (IEEE)-CS joint conference on digital libraries 2009: pp. 125-134.
[44] Sacco GM. Dynamic taxonomies and faceted search. Theory practice and experience: 1st ed. Berlin: Springer, 2009, pp. 340.
[45] Perugini S. Supporting multiple paths to objects in information hierarchies: Faceted classification, faceted search, and symbolic links. Information Processing and Management 2010; 46: pp. 22-43.

[46] Russell-Rose T and Tate T. Designing the search experience: The information architecture of discovery. 1st ed. China: Elsevier, 2013.
[47] Kaki M. Findex: Search Result Categories Help Users When Document Ranking Fails. Computer Human Interaction (CHI) 2005, Papers: Document Interaction. April 2-7th 2005, Portland, Oregon USA.
[48] Fagan JC. Usability studies of faceted browsing: A literature review. Information Technology and Libraries 2010: pp. 58-66.
[49] Kules B, Capra R, Banta M. et al. What do exploratory searchers look at in a faceted search interfaces. Joint Conference on Digital Libraries (JCDL) 2009. Proceedings of the 9th Association for Computing Machinery (ACM)/ Institute of Electricals and Electronics Engineers (IEEE)-CS joint conference on digital libraries 2009: pp. 313-322.
[50] Marchionini G. Exploratory Search: From Finding to Understanding. Communications of the Association for Computing Machinery (ACM) 2006; 49 (4): pp. 41-46.
[51] Cleverley PH. Improving Enterprise Search in the Upstream Oil and Gas Industry by Automatic Query Expansion using a Non-Probabilistic Knowledge Representation. International Journal of Applied Information Systems (IJAIS) 2012; 1(1): pp. 25-32.
[52] Hearst M. Clustering versus Faceted Categories for Information Exploration. In Communications of the Association for Computing Machinery (ACM) 2006; 49(4).
[53] Tunkelang D. Quotation. In. Russell-Rose T and Tate T. Designing the search experience: The information architecture of discovery. 1st ed. China: Elsevier, 2013, pp. 165.
[54] Antelman K, Lynema E and Pace AK. Toward a twenty first century catalog. Information Technology and Libraries 2006; 25(3).
[55] Niu X, Lown C and Hemminger BM. Log based analysis of how faceted search interact in a library catalog interface. Proceedings of Third Workshop on Human Computer Interaction and Information Retrieval 2009: pp. 62-65.
[56] Niu X and Hemminger BM. Beyond Text Querying and Ranking List: How People are searching through Faceted Catalogs in Two Library Environments. Proceedings of the 73rd Association for Information Science and Technology (ASIS&T) Annual Meeting on Navigating Streams in an Information Ecosystem 2010; 47(29)
[57] Ballard T and Blaine A. User search limiting behaviour in Online Catalogs. Comparing classic catalog use to search behaviour in next generation catalogs. New Library World 2011; 112(5/6): pp. 261-273.
[58] Kules B and Capra R. Influence of training and stage of search on gaze behaviour in a library catalog faceted search interface. Journal of the American Society for Information Science and Technology 2010; 63(1): pp. 114-138.
[59] Ren F. An unsupervised cascade learning scheme for 'cluster-theme keywords' structure extraction from scientific papers. Journal of Information Science 2014; 40(2): pp. 167-179.
[60] Chen H, Yim T, Fye, D et al. Automatic thesaurus generation for an electronic community system. Journal of the American Society for Information Science 1995; 46(3): pp. 175-193.
[61] Ding Y, Chowdhury GG and Foo S. Incorporating the results of co-word analyses to increase search variety for information retrieval. Journal of Information Science 2000; 26(6): pp. 429-452.
[62] Buzydlowski JW, White HD and Lin X. Term co-occurrence analysis as an interface for digital libraries. In: Börner K and Chen C (Eds): Visual Interfaces to Digital Libraries. 1st ed. Berlin: Springer, 2002.
[63] Shiri A. Metadata enhanced visual interfaces to digital libraries. Journal of Information Science 2008; 34(6): pp. 763-775.
[64] Olson TA. Utility of a faceted catalog for scholarly research. Library Hi Tech Emerald, 2007; 25(4): pp. 550-561.
[65] Gwizdka J. What a difference a tag cloud makes: effects of tasks and cognitive abilities on search results interface use. Information Research 2009; 14(4).
[66] Low B. Usability and contemporary user experiences in digital libraries. For metadata and Web 2.0 – Cataloguing and Indexing Group Scotland (CIGS) Seminar, 23rd February 2011, University of Edinburgh. http://www.slideshare.net/scottishlibraries/ux2-usability-and-contemporary-user-experience-in-digital-libraries (accessed January 2014).
[67] Heimerl F, Lohmann S, Lange et al. Word Cloud Explorer: Text Analytics based on Word Clouds. Proceedings of the 47th Hawaii International Conference on System Science (HICSS 2014), January 6-9th 2014. http://www.vis.uni-stuttgart.de/uploads/tx_vispublications/HICSS2014_Word_Cloud_Explorer_preprint.pdf (accessed January 2014).
[68] Halvey M and Keane MT. An assessment of tag presentation techniques. Proceedings of 16th International World Wide Web Conference (WWW) 2007. http://www2007.org/htmlposters/poster988/ (accessed January 2014).
[69] Creswell JW and Plano Clark VL. Designing and Conducting Mixed Methods Research. 2nd ed. United States of America:Sage, 2011.
[70] Stenmark D and Jadaan T. Intranet Users' Information-Seeking Behaviour: An Analysis of Longitudinal Search Log Data. In 69th Annual Meeting of the American Society for Information Science and Technology (ASIST), Austin (USA) 3-8th November 2006; 43(1).
[71] Raskin R. National Aeronautics Space Administration (NASA) Semantic Web for Earth and Environmental Terminology (SWEET) Ontology. http://sweet.jpl.nasa.gov/ontology/ (2011, accessed February 2013).
[72] Berger PL and Luckmann T. The social construction of reality. A treatise in the sociology of knowledge. 1st ed. London: Penguin, 1966.

[73] Society of Petroleum Engineers (SPE). Petrowiki based on Petroleum Engineering Handbook (PEH). http://petrowiki.org/PetroWiki (2014, accessed April 2014).
[74] Google. Google n-gram viewer. https://books.google.com/ngrams (2013, accessed September 2013).
[75] Vorhees E and Hamman D. Overview of the seventh text retrieval conference (TREC-7). National Institute of Science and Technology (NIST) Publication 500-242 1998: pp.1.
[76] Cox A and Fisher M. Expectation as a mediator of user satisfaction 2004. WWW 2004 Conference workshop (New York, May 18, 2004).
[77] Harris Z. Distributional Structure. Word 1954; 10 (23): pp. 146-162.
[78] Bartsch S. Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Cooccurrence 2004: pp. 92.
[79] Lund K, Burgess C, Atchley RA. Semantic and associative priming in high-dimensional semantic space. In. Cognitive Science Proceedings 1995: pp. 660-665. http://locutus.ucr.edu/Reprints.html (accessed October 2013).
[80] Bullinaria JA and Levy JP. Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. Behaviour Research Methods 2007; 39: pp. 510-526.
[81] Veling A and van der Weerd P. Conceptual grouping in word co-occurrence networks. Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI) 1999; 2: pp. 694-699.
[82] Vechtmova O, Roberston S, Jones S. Query Expansion with Long Span Collocates. Journal of Information Retrieval 2003; 6(2): pp. 251-273.
[83] Python website https://www.python.org/ (accessed February 2013).
[84] University of Glasgow. Information retrieval stop word list. http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words (accessed August 2013).
[85] Church K and Hanks P. Word Association Norms, Mutual Information and Lexicography. Computational Linguistics 1991; 16(1): pp. 22-29.
[86] Glaser BG and Strauss AL. The Discovery of Grounded Theory: Strategies for Qualitative Research. 7th ed. New Brunswick and London: Aldine Publishing Company, 2012.
[87] Strauss A and Corbin JA. Basics of Qualitative Research. Techniques and procedures for Developing Grounded Theory. 2nd ed. United States: Sage Publications, 1998.
[88] Dey I. Grounding Grounded Theory: Guidelines for Qualitative Inquiry. 1st ed. United Kingdom: Emerald Group, 1999.
[89] Kidwell P, Lebanon G and Cleveland WS. Visualizing incomplete and partially ranked data. Institute of Electricals and Electronics Engineers (IEEE) Transactions on Visualization and Computer Graphics 2008; 14(6): pp. 1356-1363
[90] McCallum AK. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu/ (2002, accessed November 2013).
[91] Blei D, Ng A and Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003; 3: pp. 993–1022.
[92] Balagopalan A. Topic Modelling Tool. http://code.google.com/p/topic-modeling-tool/ (2011, accessed November 2013).
[93] Kralisch A and Berendt B. Cultural determinants of search behaviour on websites. Proceedings of the International Workshop on Internationalisation of Products and Systems (IWIPS) 2004 Conference on Culture, Trust and Design Innovation. Vancouver, Canada, 8 - 10 July, 2004. http://csi.ufs.ac.za/resres/files/Kralisch(6).pdf (accessed January 2014).
[94] Marcos M, Garcia-Gavilanes R, Bataineh B et al. Cultural differences on seeking information. An eye tracking study. Computer Human Interaction (CHI) April 27 – May 2nd 2013 Paris, France. http://repositori.upf.edu/handle/10230/20943 (accessed November 2013).
[95] Sapsford R. Survey Research. 2nd ed. United Kingdom: Sage, 2007; pp. 169.
[96] Heinstrom J. Fast surfers, Broad scanners and Deep Divers – personality and information seeking behaviour. 2002 Doctoral dissertation Abo Akademi University. http://users.abo.fi/jheinstr/ (accessed June 2013).
[97] Halder S, Roy A and Chakraborty P. The influence of personality traits on information seeking behaviour of students. Malaysian Journal of Library & Information Science 2010; 15(1): pp. 41-53.
[98] Theodosiou T, Vizirianakis IS, Angelis L et al. MeSHy: Mining unanticipated PubMed information using frequency of occurrences and concurrences of MeSH terms. Journal of Biomedical Informatics 2011; 44: pp. 919-926.
[99] McCay-Peet L and Toms E. Measuring the dimensions of serendipity in digital environments. Information Research 2011; 16(3)
[100] White HD and Mayr P. Pennants from descriptors. Networked Knowledge Organizational Systems (NKOS) Workshop 2013.
[101] Spiteri LF. Word Association Testing and Thesaurus Construction. Library and Information Science Research Electronic Journal (LIBRES) 2004; 14 (2).
[102] Pisi G. Contextual Search: Issues and Challenges. Usability Symposium (USAB) 2011 Lecture Notes in Computer Science (LNCS); 7058: pp. 23-30.
[103] Beresi UC, Kim Y, Song D, Ruthven I. Why did you pick that? Visualizing relevance criteria in exploratory search. International Journal of Digital Libraries (IJDL) 2010; 11(2): pp. 59-74.
[104] Chuang J, Manning CD, Heer J. "Without the Clutter of Unimportant Words": Descriptive Keyphrases for Text Visualization. ACM Transactions on Computer-Human Interaction (TOCHI) 2012; 19(3)

[105] White RW and Roth RA. Exploratory search : beyond the query-response paradigm. 1st ed. California: Morgan & Claypool, 2009.
[106] Wilson ML. Search User Interface Design. In: Marchionini (ed). Synthesis Lectures on Information Concepts, Retrieval, and Services: Morgan & Claypool, 2011.
[107] Beavan D. Glimpses through the clouds: collocates in a new light. Digital Humanities 2008, 25th-29th June, University of Oulu.
[108] Kuo BY, Hentrich T, Good BM, Wilkinson MD. Tag Clouds for Summarizing Web Search Results. Poster Paper 16th International World Wide Web Conference May 8th-12th 2007, Banff, Alberta, Canada.
[109] Membership society and association online libraries sites. Society of Petroleum Engineers (SPE). OnePetro Digital Library Search: https://www.onepetro.org/ , American Association of Petroleum Geologists (AAPG): AAPG Datapages: http://archives.datapages.com/data/index.html (accessed January 2014).
[110] Subscription, journal online library sites and web search engines. Elsevier. ScienceDirect Digital Library Search: http://www.sciencedirect.com/, Springer. SpringerLink Digital Library Search: http://link.springer.com/, Wiley. Wiley Online Digital Library Search: http://onlinelibrary.wiley.com/, Sage: Sage Library http://library.sage.edu/, Emerald Group. Emerald Insight: http://www.emeraldinsight.com/, Google Scholar http://scholar.google.co.uk/ , Microsoft Academic Search: http://academic.research.microsoft.com/ , Stanford University: Highwire Online Literature Search: http://highwire.stanford.edu/ , Massachusetts Institute of Technology (MIT) Library: http://libraries.mit.edu/ , University of Cambridge Library: http://www.lib.cam.ac.uk/ , University of Oxford Bodleian Library (SOLO): http://www.bodleian.ox.ac.uk/ , Worldcat https://www.worldcat.org/ , Deep web http://deeperweb.com/ , Clusty http://www.clusty.com/ (all accessed January 2014).
[111] Enterprise search and text analytics sites and articles. Microsoft: Enterprise Search Case Study Online Article: http://www.microsoft.com/casestudies/Microsoft-FAST-Search-Server-2010-For-Sharepoint/International-Monetary-Fund/Economists-Quickly-Find-Information-Using-Enterprise-Search/4000011274 Google: Enterprise Search Case Study Online Articles: http://www.computerweekly.com/feature/Case-Study-How-BG-Group-made-Google-work-internally and http://www.slideshare.net/searchtechnologies/a-gsa-case-study-manufacturing-industry. Enterprise Search Europe 2014 Conference Presentations. http://www.enterprisesearcheurope.com/2014/ . Findwise Enterprise Search Case Study SlideShare Channel: http://www.slideshare.net/findwise , Butler Analytics: 2014 Online Article Text Analytics Platforms: http://butleranalytics.com/20-text-analytics-platforms/ , Expert System http://www.expertsystem.net/ (all accessed March 2014).
[112] Government online library sites. Carmarthenshire (Wales, UK) Local Government Library: http://www.carmarthenshire.gov.uk/english/education/libraries/Pages/Home.aspx, United States Library of Congress Online Public Access Catalog (OPAC): http://www.loc.gov/library/libarch-digital.html (both accessed March 2014).
[113] Rasmus DW. The Serendipity Economy. Harvard Business Review (HBR) Online Article http://blogs.hbr.org/2013/08/how-it-professionals-can-embrace-the-serendipity/ (2013, accessed May 2014).