Creating sparks: comparing search results using discriminatory search term word co-occurrence to facilitate serendipity in the enterprise.

Paul Hugh Cleverley

Robert Gordon University, Garthdee Road, Aberdeen, United Kingdom AB10 7QB. p.h.cleverley@rgu.ac.uk

Simon Burnett

Robert Gordon University, Garthdee Road, Aberdeen, United Kingdom AB10 7QB. s.burnett@rgu.ac.uk

Abstract: *Categories or tags that appear in faceted search interfaces which are representative of an information item, rarely convey unexpected or non-obvious associated concepts buried within search results. No prior research has been identified which assesses the usefulness of discriminative search term word co-occurrence to generate facets to act as catalysts to facilitate insightful and serendipitous encounters during exploratory search. In this study, fifty three scientists from two organizations interacted with semi-interactive stimuli, 74% expressing a large/moderate desire to use such techniques within their workplace. Preferences were shown for certain algorithms and colour coding. Insightful and serendipitous encounters were identified. These techniques appear to offer a significant improvement over existing approaches used within the study organizations, providing further evidence that insightful and serendipitous encounters can be facilitated in the search user interface. This research has implications for organizational learning, knowledge discovery and exploratory search interface design.*

Keywords: *Enterprise search, digital library, exploratory search, serendipity, text analytics, information discovery*

Introduction

Categories or tags that are *representative* of information items rarely convey the unexpected. This research explores the immediately matched search term word co-occurrences that exist within items in search results, as these word co-occurrences are not representative of the item as a whole (but are representative of the *search context*). Word co-occurrence techniques utilize the words that appear close to the target search terms in the body text of a document as independent associations that may represent a meaningful real world association. The use of such an approach may generate more unexpected terms (prompts) to interact with and stimulate insightful and serendipitous encounters.

The use of text analytics techniques can automate the *tagging* process and create a form of *dynamic* natural language indexing, where the terms presented to the searcher (for any given information item) change as the search terms change. Furthermore, the research examines how comparing these word co-occurrences by differing contexts and displaying the resulting *discriminatory* co-occurring terms can surface patterns that may otherwise be obscured, offering potential insightful and/or serendipitous encounters.

From the literature an initial inter-disciplinary theoretical model is developed. The research uses a stimulus to provoke responses from participants, using a controlled vocabulary to colour code the subsequently dynamically generated word associations. The interactions are directly observed and quantitative data is gathered using questionnaires. From the resulting analysis, the research aims to understand the extent to which serendipitous encounters can be facilitated by these approaches and why. The theoretical model is revisited in light of the empirical findings generated by the study and concomitant refinements made.

Research Questions and Aims

The research addresses the following questions:

1. Can comparison of search results using discriminatory search term word co-occurrence facilitate serendipity in the enterprise?
2. How does this happen?
3. When would these techniques be of value to an organization?

It was hypothesized that certain text analytic methods (discriminatory search term word co-occurrence) could surface unexpected terms buried in search results which have the potential to facilitate serendipity to a moderate/large extent. It was also hypothesized that these text analytic methods may enhance existing search tools in the enterprise with respect to exploratory searching. An initial strawman theoretical model (Figure 1) was developed from the literature

(Cleverley & Burnett 2014, Makri & Blandford 2012, Toms & McCay-Peet 2009, Marchionini 2006). This theoretical model is revised based on the findings from the study.

The aims of this study were to determine the extent to which discriminatory search term word co-occurrence techniques can facilitate serendipity in the enterprise.

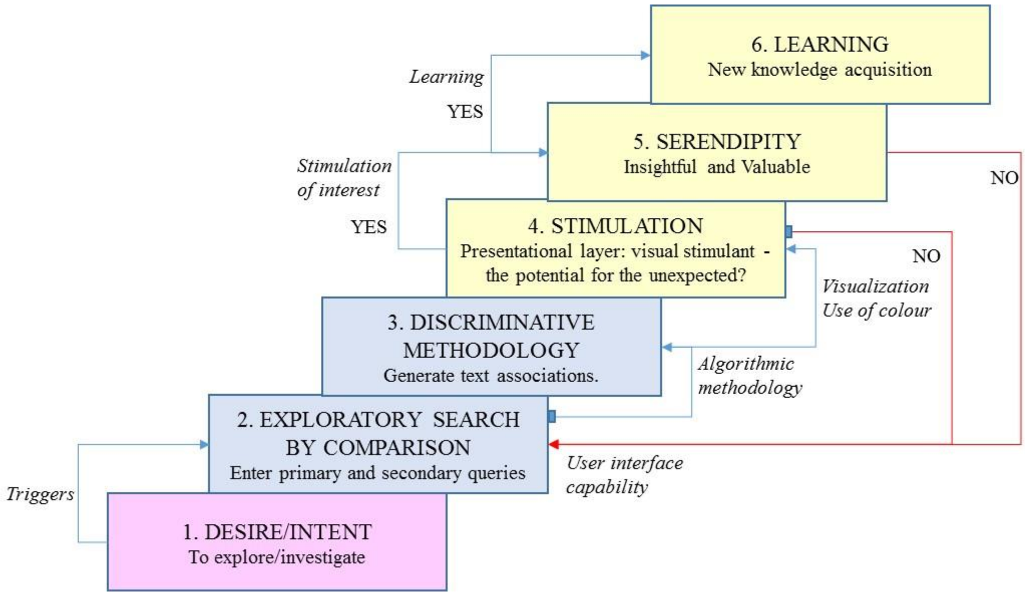

6. LEARNING
New knowledge acquisition
Learning
YES
5. SERENDIPITY
Insightful and Valuable
NO
Stimulation of interest
YES
4. STIMULATION
Presentational layer: visual stimulant - the potential for the unexpected?
NO
Visualization
Use of colour
3. DISCRIMINATIVE METHODOLOGY
Generate text associations.
Algorithmic methodology
2. EXPLORATORY SEARCH BY COMPARISON
Enter primary and secondary queries
User interface capability
Triggers
1. DESIRE/INTENT
To explore/investigate

Figure 1 – Initial strawman theoretical model for discriminatory search term word co-occurrence

Research value

No prior research has been identified that studies these particular serendipity-by-comparison techniques in an enterprise. It is anticipated that researchers in a number of academic disciplines (information technology, information management, knowledge management and organizational learning) could utilize some of the theoretical models and empirical findings within their own studies.

Developing an understanding of how enterprise search can facilitate serendipity may help transform beliefs and practices in the organization. In economic terms, enterprise search technology is traditionally seen by organizations as an efficiency saving measure, focusing on reducing the time staff spend looking for information (White 2012, Feldman & Sherman 2011). This is often reflected in the approach and level of investment made by organizations towards this area. There has been little attention paid within the enterprise (perhaps because it is less tangible and more challenging to measure) to the extent to which enterprise search can facilitate insightful and serendipitous encounters, sparking creativity, innovation and ultimately generating business value. Rasmus (2013) proposes the *serendipity economy*, a space where unanticipated interactions occur that may produce business value which are impossible to predict. If an enterprise can develop a greater capability to facilitate serendipity within its search interfaces than the *status quo*, it may gain business performance benefits.

Literature Review

Information landscape

Enterprise information volumes are doubling every two years (Gantz & Reinsell 2011). With searchers entering small amounts of query information into user interfaces (Taghavi *et al.* 2011) searching ever larger corpus sizes, it is unsurprising that queries may return hundreds if not thousands of search results. Indeed, there is survey evidence that enterprise staff may have as much difficulty finding information today as they did ten years ago (Mindmetre 2011).

Serendipity

Innovation or creativity sparked by an unexpected seemingly random event is often called serendipity. Within organizations, the discovery of innovations and business opportunities is often serendipitous (de Rond & Morley 2010, Ghiselin 2010, Friedman 2010, Denrell *et al.* 2003, Meyer & Skak 2002). Within the context of this research, serendipity is defined as the phenomenon of fortuitous unexpected information discovery.

Serendipity may be an inevitable consequence of immersion within information rich environments (McCay-Peet & Toms 2011) making hitherto unforeseen connections. Studies of unanticipated epiphany have shown that a prerequisite to serendipity is sagacity, a *prepared mind* (Rubin *et al.* 2011, McBirnie 2008, Foster & Ford 2003). Serendipity as a phenomenon is unlikely to be controllable; however, developing a *capability* that may lead to more opportunities for serendipitous encounters is considered plausible. Creativity often requires a diverse range of inputs (Daveport & Prusak 2000). Some researchers believe there is a general human propensity for the *unexpected* (observer dependent), both as an agent and recipient (Dessalles 2009). This premise postulates that human cognition is highly sensitive to any discrepancy in complexity.

In the *physical world*, the concept of facilitating serendipity in the enterprise through connectivity is well established. These include facilitating co-presence knowledge sharing, innovation and serendipity through facility design (Appel-Meulenbroek 2010), internal and external social networks (Rasmus 2013) and traditional physical libraries which allow passive browsing of *the stacks*, enabling systematic serendipity (Rice 1988, Smith 1964).

In the *virtual world* facilitating serendipitous connections has attracted significant attention. Knowledge organization methods applied to physical libraries, can also be applied to digital libraries. Hyperlinked networks (part of the Internet phenomena) are methods which can foster serendipity through browsing. Making use of historical crowdsourcing data (collaborative filtering) to suggest and recommend information based on usage data has become a popular method to stimulate serendipitous experiences (Zhang *et al.* 2012). Social media platforms such as microblogging (e.g. Twitter) have also been shown to facilitate serendipitous encounters enabling diverse and unexpected connections to be presented and made (Martin & Quan-Haase 2014). With the ability to search and access vast amounts of information, navigating from one place to another, some declare *"The Internet is the greatest serendipity engine in the history of culture"* (Thompson 2006). These methods are not without their criticism. Both researchers and those in the media raise concerns that algorithms are increasingly creating *filter bubbles* and reducing true serendipity. As put by Leslie (2012) *"The Internet has become so good at meeting our desires that we spend less time discovering new ones"*.

Just as in the physical world, challenges exist to facilitate serendipity in a digital environments within the enterprise, whilst mitigating the potential for distraction (Siefring *et al.* 2012, Wilson & Schraefel 2008).

There are likely to be degrees of serendipity, ranging from a chance encounter whilst undertaking a totally unintended intention, through to a chance encounter whilst undertaking a related intention. For example, a geoscientist 'bumping into' an engineer from another company during lunch at an oil and gas conference (after speaking with many of the attendees) which results in a serendipitous encounter, is quite different in nature to a situation if the same encounter happened in a supermarket. There may be a *peculiarity spectrum* pertaining to the events leading to the chance encounter, which in turn may be related to information seeking strategies. Both however, may be classed as unexpected, surprising and valuable events. Judging whether an encounter is serendipitous or not, is likely to be subjective. The serendipity space model (Makri & Blandford 2012) uses three elements; unexpectedness, insightfulness and value. The model states that all elements must be present for an encounter to be serendipitous, but recognizes a spectrum of encounters, e.g. from *somewhat unexpected* to *very unexpected*.

It is possible that some discoveries which are surprising in themselves, have been misclassified as serendipitous. For example, Wallmart discovered through integrating disparate data, that its Pop-tart product sales increase seven-fold when there is a hurricane warning. This was reported in the media as serendipitous (Hulme 2012). This pattern was discovered whilst purposefully seeking product sales patterns during hurricane warnings (a technique in the wider field of predictive analytics). Whilst valuable for the retailer, it could be argued that this is not a serendipitous encounter as the *intent* was very specific and in that context, the encounter not really unexpected.

Knowledge organization

Knowledge organization systems such as controlled vocabularies (taxonomies and thesauri) allow enterprises to index (tag) their information to aid subsequent retrieval (Cleverley 2012, Soergel 2009). This retrieval may be through search or through browsing of the tags, supporting the exploratory search process. It is likely the same *peculiarity spectrum* that exists in the physical world that exists in the virtual world, related to the intent of the information seeker when they started to browse and the subsequent information encountered when browsing.

Natural language indexing allows any term in the document to be used for tagging, differing from the controlled list approach. Free text indexing takes natural language indexing a step further to allow any term to be used for tagging, whether it is the document (or corpus) or not.

The process of tagging can be a manual process performed by the authors of the information, or provided as a service (e.g. from a corporate library). Classification schemes can also be driven by a community, created without any central control, through cooperation and re-use - often termed folksonomies. Tagging can also be achieved through automated processes using statistical and/or linguistic rules, which classify information to a controlled vocabulary or are completely natural language data driven (Zamir & Etzioni 1999).

These approaches are typically *static*, in that the displayed tags of any given item do not change once assigned. For example, a specific document tagged as a 'Hydrocarbon prospectivity report' and 'Trinidad', will only ever show these two tags within a faceted search menu to represent its existence in the results set, regardless of the search terms used. Even if some form of thesauri or ontology enriches these tags (e.g. the ontology contains knowledge that this type of report is *used by* geoscientists), they are none the less fixed tags that represent *the whole*. This presents limitations for browsing. For example if the searcher makes a query on 'sparse spike inversion', it is possible out of hundreds or thousands of search results, the report in question may be the only one with the term 'shoreface sands' co-occurring with the search phrase 'sparse spike inversion' in its text. That unusual or discriminatory fact will, however, remain hidden from the searcher who will only see the tags that represent the document as a whole in the faceted search menu.

Information retrieval and exploratory search

Through some of the seminal research in this area, search goals have been grouped into three categories termed *navigational* (locate a website), *informational* (locate information) and *transactional* (perform an activity) (Broder 2002). This was further refined in the *informational* category (Rose & Levison 2004) to include both closed (e.g. What date is the 2018 football world cup?) and open questions of unconstrained depth (e.g. What is the relationship between bicycles and road accidents?).

Search activity has also been grouped in two broad categories, lookup (known items) and exploratory (where the question is not fully formed in the mind of the searcher and/or where search terms cannot be specified in advance) (Marchionini 2006). Many information needs for *known items* may be fulfilled through a simple search box in combination with a good set of ranked search results. This *Google type* of search, which incorporates usage, authority and currency in its ranking (Brin & Page 1998), also relies on a keyword tagging strategy similar to library catalogues. If the item(s) sought are retrieved on the first page(s) or towards the top of the ranking, information goals may be met. This type of search activity is the most common by volume, focusing on information precision. Clusters of this type of activity have been identified from search logs termed *hit and run, popular and focused* (Wolfram *et al.* 2009).

The need for greater involvement from the searcher has been recognized with the emergence of the human computer information retrieval discipline (Marchionini 2006) combining the fields of human computer interaction and information retrieval. Exploratory search activity, to *learn or investigate* involves a browsing component and has also been identified from search logs, termed *long and varied* (Wolfram et al. 2009). Exploratory search tasks may comprise 8-27% of all search usage (Chapman *et al.* 2013, Stenmark 2008) and may support some high value organizational needs (e.g. learning and innovation). A study of the *long and varied* clusters in search logs has identified a proportion of use (36%) that may simply represent struggling, rather than exploratory search (Hassan *et al.* 2014). This highlights the difficulty in researching search behaviour (Wilson 1999) from single research methods.

Exploratory search has been sub-divided into a number of modes including analyze>*compare*>evaluate>knowledge acquisition (Russell-Rose *&* Tate 2011, Morville & Callendar 2010, Marchionini 2006). Research in exploratory search user interface design is of considerable and ongoing interest, particularly using text analytics and graphical representations (Sarrafzadeh *et al.* 2014, Yogev 2014, Haun & Nurnberger 2013, Nitsche & Nurnberger 2013, Ruotsalo *et al.* 2013, Nunez *et al.* 2011, Yang & Wagner 2010, Kules et al. 2009, Kules & Schneiderman 2010, 2007). Topic trends and entities within documents have been visualized through time and space (Reinanda *et al.* 2013, Hoffart *et al.* 2011, Krestel *et al.* 2011, Shi *et al.* 2010) and other data driven contexts (Khalili *et al.* 2014). There appears to be little research on discriminatory based techniques applied to search contexts that could be controlled by the end user of the system. For example, to explore the results in a library that contain the term waterflooding (an enhanced oil recovery technique), comparing what is different in the search term co-occurrences for the documents that mention 'waterflooding AND dolomite', compared to the documents that mention 'waterflooding AND limestone' compared to the documents that mention 'waterflooding AND sandstone'. It may be possible that for the searcher who *may not have a very specific prior notions*, unexpected terms are surfaced leading to surprising results.

Empowering people to search and learn has been identified as both an opportunity and challenge in information retrieval (Allan *et al.* 2012, pg. 9): *"helping people to achieve higher levels of learning through the provision of more sophisticated, integrative and diverse search environments… make tools that will lead to meaningful outcomes to motivate adoption"*. Increasing the propensity of the search user interface to facilitate serendipity has been studied to some extent (Makri *et al.* 2014, Alexander *et al.* 2014, Andre *et al.* 2009, Toms & McCay-Peet 2009). This recognizes that many searchers may have an intent personified by, *"show me something I don't know already"* (Nolan 2008, pg. 38). Allan *et al.* (2012, pg.11) describes how "*these tools are likely to interrupt and disrupt a comfortable searching style"*, a belief taken further by some researchers, that current Internet search tools have made enterprise information seekers *lazy* (Sweeny 2011), unable to create complex searches and who rarely explore past the first few search results.

Browsing in search interfaces

Browsing has been shown to support creativity, whether the intent is purposive, capricious or exploratory in nature (Bawden 1986). Compared to the Internet, an enterprise has information that is not always web authored, so it lacks the network of hyperlinks for browsing from topic to topic. Faceted search is a technique which also allows browsing, supporting exploratory search; presenting topics within the user interface inviting further human interaction to filter results. Faceted search has been shown to improve search effectiveness (Fagan 2010). Terms within facets typically come from underlying controlled vocabularies, with information manually tagged or automatically classified to categories within that vocabulary (La Barre 2010). Facet values for browsing purposes can also be data driven, generated automatically (clustered) from text. Clustering can be applied to the entire document texts within search results (Yogev 2014, Scaiella *et al.* 2012, Palmer *et al.* 2001) or a matched context window within the document, using words that co-occur with the search terms (Kaizer & Hodge 2005). Challenges identified for faceted search include "*how many facets should be displayed in a given context in what order and, most importantly, how should the most relevant facets be identified*" (Teevan *et al.* 2008, pg.2).

A useful characteristic of word co-occurrence (a technique in the broader field of text analytics) is its ability to create an associative network to surface potentially unexpected contexts. In particular, allowing the searcher to visualize what is common (representative) and what is different (discriminative). Using rich and non-obvious word associations has the potential to stimulate serendipity (Cleverley & Burnett 2014). Studies of search user interfaces indicate word co-occurrence filters may aid discovery (Gwizdha 2009, Olsen 2007) other studies showing no use at all (Low 2011).

Presentation

The Gestalt laws (Wertheimer 1923) have been applied to search user interface design (Michalski 2014, Change *et al.* 2002), including the use of colour to group information. People are attracted by visually salient stimulus, a concept often used in tag clouds (e.g. text size, centrality, hue and lightness) to highlight patterns which may otherwise remain obscured (O'Donnell 2011, Stasko *et al.* 2008). Use of colour to group categories of similar things has been used in traditional faceted search (Hearst & Stoica 2009) and infographics (McCandless 2012). Novel graphical visualizations have been used to stimulate serendipitous information discovery (Thudt *et al.* 2012) however, searchers may prefer the utility of simple lists instead of word cloud type displays (Heimerl *et al.* 2014, Rivadeneira *et al.* 2007), people finding lists faster to scan and navigate than novel spatial representations (Halvey & Keane 2007).

In summary, serendipity as a phenomenon is attracting increasing attention within organizations because of the extraordinary business value it can generate. Numerous approaches based on text analytics are emerging to complement (and perhaps challenge) the traditional *search box and results list* and the manually created *browse-able menu structures* used by digital libraries. It is believed that no empirical studies exist which observe how practicing scientists interact with discriminatory comparison based approaches using text analytics, when applied to enterprise content. This study aims to address this gap in the literature and act as a catalyst for further research in this area.

Method

Philosophy, strategy and design

A positivist (deductive) and constructivist (inductive) approach was adopted for the research. This was designed to address confirmatory and exploratory questions, enabling the triangulation of the insights, shared understandings and differences of views in the area investigated. Statistically significant preferences were identified. A pragmatic and dialectic (Berniker & McNabb 2006) stance was taken to gain insight and explore differences, with an approach based on grounded theory (Strauss & Corbin 1998) to thematically map nuances and comments.

A cross sectional mixed methods sequential, convergent and multi-phase design was chosen (Figure 2). This allowed participant data from two organizations to be integrated, the results of the first organization influencing the questions to the second organization and the generation of the visual stimulus (semi-interactive prompt sheet). The quantitative data would be used to test the internal and external validity, which is integrated with the qualitative data to triangulate its internal trustworthiness and external transferability.

Data collection and sampling

Focus groups (Marshall & Rossman 2006, Morgan 1997) were chosen to collect the data for the following reasons:

- Focus groups allow spontaneous interactions between participants, they can talk to each other ask questions, state opinions and share experiences. They allow participants to clarify their own understandings and differences with one another.
- The visual stimulus is only semi-interactive, so participants would be required to do a fair amount of conceptualization. It is believed this is more effective to do within a group setting (Furnham 2000).
- The potential for focus groups to develop unanticipated arguments is of significant interest for this research.
- The amount of time and access the researchers had with practicing business professionals was limited. A focus group is a way to elicit a diverse range of views in a short space of time.

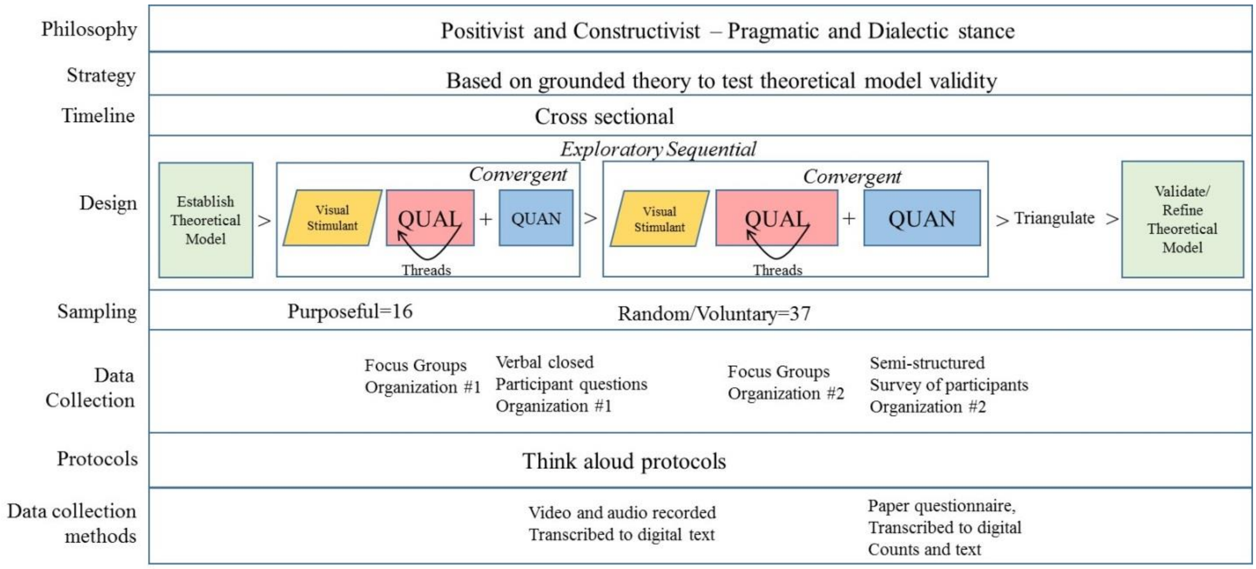

Philosophy
Positivist and Constructivist – Pragmatic and Dialectic stance
Strategy
Based on grounded theory to test theoretical model validity
Timeline
Cross sectional
Design
Exploratory Sequential
Establish Theoretical Model
Convergent
Visual Stimulant
QUAL
+
QUAN
Threads
Convergent
Visual Stimulant
QUAL
+
QUAN
Threads
Triangulate
Validate/ Refine Theoretical Model
Sampling
Purposeful=16
Random/Voluntary=37
Data Collection
Focus Groups Organization #1
Verbal closed Participant questions Organization #1
Focus Groups Organization #2
Semi-structured Survey of participants Organization #2
Protocols
Think aloud protocols
Data collection methods
Video and audio recorded Transcribed to digital text
Paper questionnaire, Transcribed to digital Counts and text

Figure 2 – The research philosophy and design

However, focus groups have some general limitations. They may be hard to moderate and keep on research topic, however with a large visual stimulus used in the focus group, it was thought this was unlikely to be a significant factor. Participants may feel uncomfortable sharing their views in a group environment. For this research topic (exploratory search) and participant background (all geoscientists in the same company), it was believed that personal emotion and organizational hierarchy would probably not be a major constraining factor.

Initial pilot testing of the visualization with colleagues indicated there was some confusion as to how the terms had been created and therefore what could be inferred. All focus groups therefore had an introductory presentation on the purpose of the research and the word association provenance. It is accepted that observer-expectancy effects may be present, where the researcher's cognitive bias could have caused them to unconsciously lead the participants towards an expected result. Focus groups require moderation (the researchers), so are subject to moderator bias. This was mitigated to a certain extent by having a control focus group which had no interaction with the moderator at discussion time. A questionnaire at the end of each focus group ascertained if the control group had statistically significant differences from other groups which were moderated.

The focus groups were held at the respective company premises and video recorded. Focus groups are by their very nature artificial environments, so participants were being asked to discuss their preferences and thoughts on a semi-interactive stimulus outside the natural environment in which it would be used. This is very different to a case study. This limitation is acknowledged, but none the less it is a useful study to stimulate insights on the phenomenon of serendipity in enterprises and provide input into further research for exploratory search interfaces.

The transferability of focus group data can be an issue as samples are small and each focus group has its own dynamics (Vicsek 2010). By purposefully sampling quite different organizations within the upstream oil and gas industry (a large multinational oil and gas operator and a small geoscience information provider in different countries) and holding multiple focus groups within each organization, the triangulation of results would help identify possible transferable themes. Using multiple methods with a series of closed questions asked at the end of each focus group enabled specific comparison of quantitative data between intra- and inter-organizational focus groups to identify statistically significant findings. The use of a convergence coding matrix (Farmer *et al.* 2006) to combine all of the data from these multiple methods is used to integrate and synthesize results to deliver more robust findings.

Truly random sampling is difficult within an organization. Typically what is sampled becomes a convenience based sample. Voluntary response bias is a limitation of many qualitative research methods, as the process of turning up for a focus group or answering a questionnaire is a form of self-selection. This was mitigated to a certain extent as the session with organization #2 was a regular external 'lunch and learn' slot in their timetables, so people who attended were not those who necessarily had a personal interest in the subject of exploratory search.

A total of seven focus groups was used from two organizations to collect the data, with no more than ten participants in any one group (UofN 2013), a total of fifty three scientists sampled. A visual semi-interactive stimulus was used, researchers acted as moderators, steering the topic of conversation and answering questions. Organization #1 is a large multinational oil and gas operator, the focus group was with staff whose first language was not English. The country location cannot be provided as this may lead to the recognition of the organization which is to remain anonymous.

Organization #2 is a geoscience consultancy in the United Kingdom, with approximately 100 staff in a single location. The focus groups were held at the respective company premises, video recorded and conversations transcribed and entered into the Nvivo software for qualitative data analysis (coding).

In organization #1, geophysicists were purposefully sampled by the researchers, targeting different geophysical departments. Sixteen participants attended from a list of twenty (four staff were unable to make the timetable). The sample consisted of fourteen men and two women (a strong male bias), all with twenty to thirty (or more) years of experience.

Groups of between two and nine staff were introduced to the visualizations created by this study and their interactions video recorded (direct observation). At the end of the sessions, participants were asked simple verbal closed questions to collect some basic quantitative data on opinions about the ability of the techniques to facilitate serendipity.

In organization #2, the point of contact for the researchers invited all the geologists on-site, so the researchers did not influence who received an invite. Thirty seven staff participated, fifteen men and twenty two women (a more balanced gender sample than organization #1), with many new graduates and young staff. They were divided roughly into two equal groups in the morning and afternoon. Each group was further broken down into two equal groups on either side of the room, each with access to the semi-interactive visual stimuli.

All participants were asked to take a few minutes to complete an anonymous paper based questionnaire before leaving the room. This allowed some time to reflect (incubation) with the thoughts fresh in their mind. This questionnaire contained closed questions using five point psychometric Likert scales (Colman *et al.* 1997) to gather quantitative data on participant opinions and space was also provided in the questionnaire for participant comments for each question.

Focus group format

In an initial ten minute presentation, participants were informed of the research questions. Word association as a technique was also explained so participants would have a better understanding of the provenance of word associations.

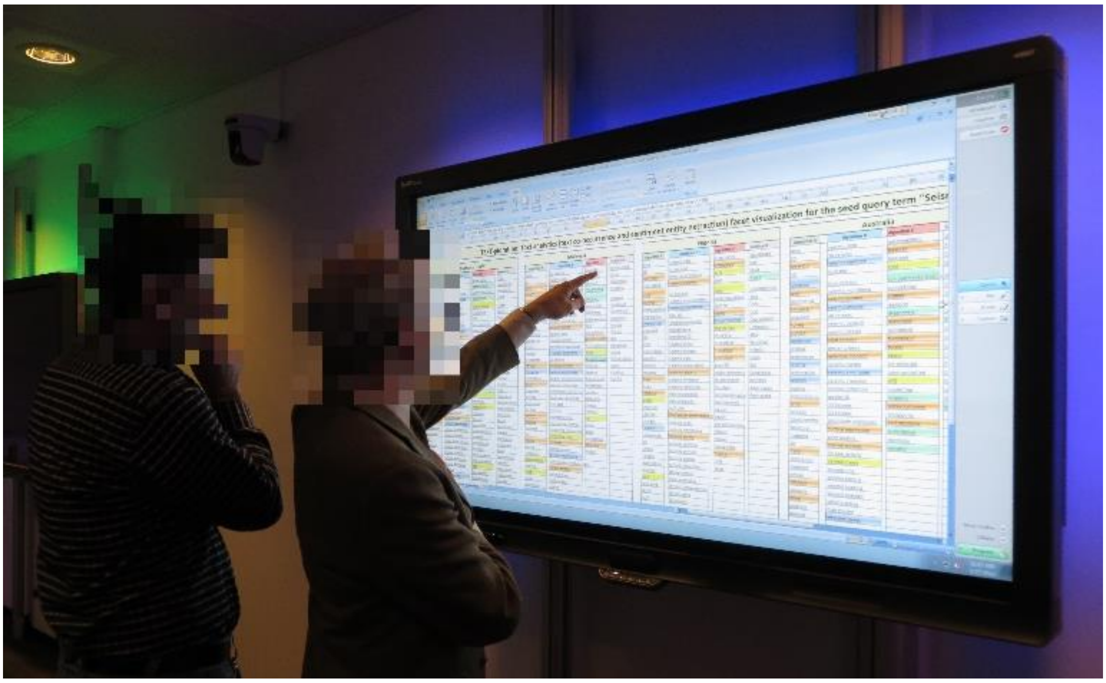


Figure 3 – Visual stimulus example of a focus group in organization #1 (faces pixelated for anonymity)

The specific task given to each focus group was to identify terms of interest in the semi-interactive visualizations and using a technique called think-aloud protocols (Beresi *et al.* 2011), state what they were thinking and why, stimulating debate within the group.

Each session lasted between forty five minutes to an hour. In organization #1 the visualization was presented on large touchscreens (Figure 3). This enabled the participants to *touch* associations of interest to them on the screen, drilling down to the individual documents in which that association exists. In organization #2 a visualization was presented as a large poster (Figure 4), as a touchscreen was not available. A tablet device was available with the visualization, should any participants identify something of interest on the poster that they wished to drill down into the actual documents.

Figure 4 – Example of a focus group in organization #2, the tablet is shown on the desk in the foreground

Association Generation Methodology

Visualizations were created as a visual stimuli for the focus groups. To ensure a particular source was not biasing results, different society membership information sources were used for each organization.

In organization #1, a digital library corpus was used from the Society of Petroleum Engineers (SPE) consisting of over 70,000 papers, relevant to the field of enquiry, simulating a subset of an enterprise corpus. The risk of biasing serendipitous experiences simply through the introduction of entirely *new* content was mitigated, as the SPE is a public domain resource which is already used by the scientists who are the subjects of the research. The Search Transaction Logs (STL) of organization #1 were briefly analyzed to gather representative exploratory technical search terms.

Word co-occurrence is used on the basis that words that appear in proximity to other words in texts share some meaning (Harris 1954). A statistical language model is created by assigning probabilities to sequences of words based on their frequency of occurrence in corpora. After a brief review of literature, first order n-grams (Bird *et al.* 2009) were chosen as they were capable of delivering a diverse set of associations with respect to the number of words, frequency of occurrence and discriminatory capability. Other co-occurrence algorithms may also deliver the same elements of diversity (for example Topic modeling), but it was not in the research scope to compare algorithms.

The primary (e.g. seismic) and secondary (e.g. Gulf of Mexico, Malaysia, Nigeria) exploratory searches were used as seed queries within the sample corpus using a word co-occurrence window ranging between fifty to sixteen words respectively (Bullinaria & Levy 2007, Vechtmova *et al.* 2003, Veling & van der Weerd 1999, Lund *et al.* 1995). The Python programming language was used to create three n-gram algorithms from the information source using the queries given and a simple sentiment entity extraction list was used for Algorithm D (Minqing & Lio 2004). Common words were filtered out using a stop word list (UofG 2013). The four lists were as follows:

- List A - Unigram (e.g. list of single terms) ranked by descending frequency
- List B - Bigram ranked by descending frequency
- List C – Discriminatory unigrams for each secondary query
- List D - An single word entity extraction of pre-defined problem /sentiment terms

Algorithms A and B effectively formed a control, as these were common words (associations), so thought unlikely to generate anything unexpected. This was evaluated with the focus groups. Algorithm C warrants more description, as it generates the discriminant associations. The key point is the discriminatory element will always depend on what secondary queries have been entered as they define the ‘collection’ of results. For the primary search term(s) P, the secondary search term(s) are (S_1, S_2, $S_3 \ldots S_{n)}$ where n is the number of secondary search terms chosen. A valid context match for a secondary term, is where a document contains both P and S within a fifty word window in the text (MW=50). For those matches, a unigram of terms (t) is generated from a sixteen word window (CW=16) around the secondary term(s), creating a co-occurrence term vector for each respective secondary search term. It follows that each secondary search term will have its own co-occurrence vector given by:

$SC_n= \{t_1, t_2, t_3, ..t_n\}$.

The universe (μ) is defined as the union of all term co-occurrences for all secondary queries:

$\mu = \{\{SC_1\} \cup \{SC_2\} \cup \{SC_3\} \cup \{SC_n\}\}$.

The discriminant terms (DS_n) for each secondary query (for example SC_1) is therefore the absolute set complement:

$DS_{1\,=} \{\{SC_2\} \cup \{SC_3\} \cup \{SC_n\}\}' \equiv SC_1 \backslash \{\{SC_2\} \cup \{SC_3\} \cup \{SC_{n\}}\} = \{x \in SC_1 \mid x \notin \{\{SC_2\} \cup \{SC_3\} \cup \{SC_n\}\}\}$

For example using the SPE collection;
If P = ‘seismic’ and S_1 = ‘gulf of mexico’, S_2 = ‘malaysia’, S_3 = ‘nigeria’, S_4 = ‘australia’ and S_5 = ‘canada’:
$DS_{1\,=}$ {‘attenuation’, ‘backscatter’, ‘bright-spot’, ..} is the set of terms that *only* occurs with P and S_1 (MW=50, CW=16)

In organization #2 a digital corpus was used from Geological Society of London (GSL) and American Geological Institute (AGI) who gave permission for this research. There were approximately 15,000 report abstracts. The algorithms used were dependent on the results from the focus group within organization #1 (see Results). Subsequently, only Algorithm C was used for organization #2 and the number of terms reduced from fifty to thirty. No STL data were available for organization #2, search terms were chosen based on the nature of the organization. A consideration was ensuring the search terms occurred enough times in the corpus to generate useful results. A primary search term of ‘carbonate’ was chosen, secondary search terms were geological periods ‘Triassic’, ‘Jurassic’, ‘Cretaceous’ etc.

Colour coding by category

After a review of suitable ontologies (using semantic web search engines), a derivation of the Semantic Web for Earth and Environmental Terminology (SWEET) was chosen (Figure 5) (Raskin 2011) to colour code co-occurring terms.

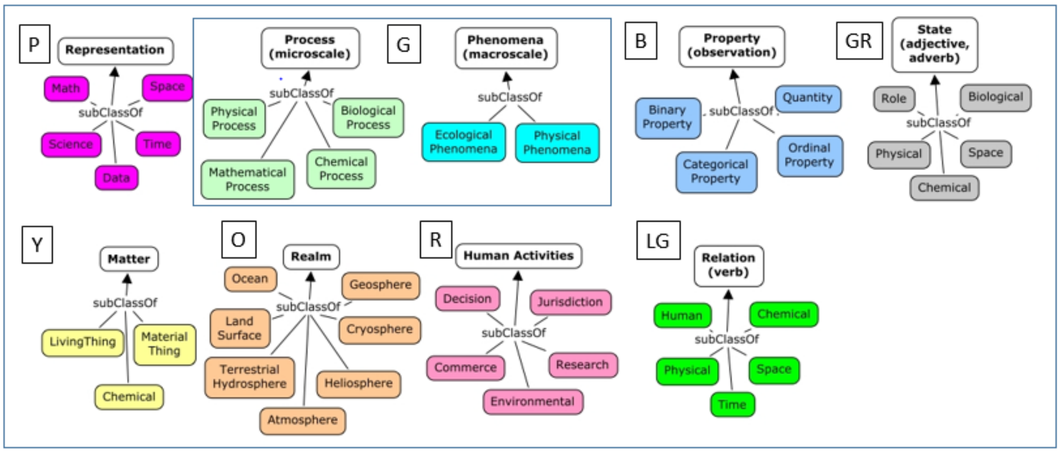


Figure 5 –SWEET Ontology (top level only) used to colour code word co-occurrences.

This suited the subject matter in question and the simplistic granularity required for the research to explore the use of colour. The SWEET ontology was supplemented by additional lists of geographical place names (Purple=PU) from sources such as http://geonames.usgs.gov/ that allowed terms which were geographical entities to be grouped with the same colour. To enable identification in monochrome for this paper, letter codes are assigned (e.g. Y = Yellow = Matter). Named Entity Recognition (NER) is the term given to automatically identifying units of information in text, for example, people, organizations and geographical entities (Nadeau & Sekine 2007). Various techniques can be used for NER including common word/phrase ending, word co-occurrence patterns, list lookups and dictionaries. For this research a list look-up was used as it delivered the required colours but could easily be expanded to include more sophisticated methods and linking to other data entities.

The four algorithms used for organization #1, as they apply to the primary search term *seismic* and secondary search term *Gulf of Mexico*, are shown in Table 1, colour coded to the ontology presented in Figure 5 (in this example the focus is only on O, B, G and Y). The discriminative term associations may have the potential to surface unexpected associations (based on the context) as determined by the observer. So the term associations need not necessarily be unusual words in themselves, the context itself could be unexpected or intriguing.

Table 1 – Example for colour coding word co-occurrence facets for the search query seismic AND Gulf of Mexico

List A	***List B***	***List C***	***List D***
Data	Seismic data	Aspectrally	Anomalies
3D	3D seismic	Attenuation [G]	Barriers
Reservoir [O]	Time-lapse seismic	Autocorrelation	Difficult
Well [O]	Seismic amplitude [B]	Backscatter [G]	Erroneous
Time-lapse	Seismic surveys	Bottom-cable [Y]	Good
Amplitude [B]	Seismic reflection	Bright-spot	Hazard
Interpretation	4D seismic	CDP	Important
Velocity [B]	Seismic response	Deconvolution	Issue
...	…	…	…

Representation – semi-interactive visualization

There are numerous ways to visually represent associations, for example node-link diagrams, Euler diagrams, scatterplots, ribbons and tree-maps (Streit *et al.* 2012). A tabular correlation (from left to right) was chosen as it was deemed the most efficient way to present a large number of non-hierarchical terms, some of which were contextually discriminant. The first fifty associations were chosen to increase the amount typically displayed in facets (e.g. most faceted search interfaces used by the enterprise display five by default which can be expanded to twenty). Despite displaying a large number of terms, as the searcher would be scanning (with no need to remember previous sequences) the cognitive load would be low, an important design criteria. Algorithms A and B were based on descending frequency. It was decided to rank Algorithms C and D alphabetically (Beall 2007) as it was felt for these lists the frequency of occurrence of the term(s) may be less relevant, the juxtaposition of terms of potential interest. The visualization used for organizations #1 and #2 is shown in Figures 6 and 7 respectively.

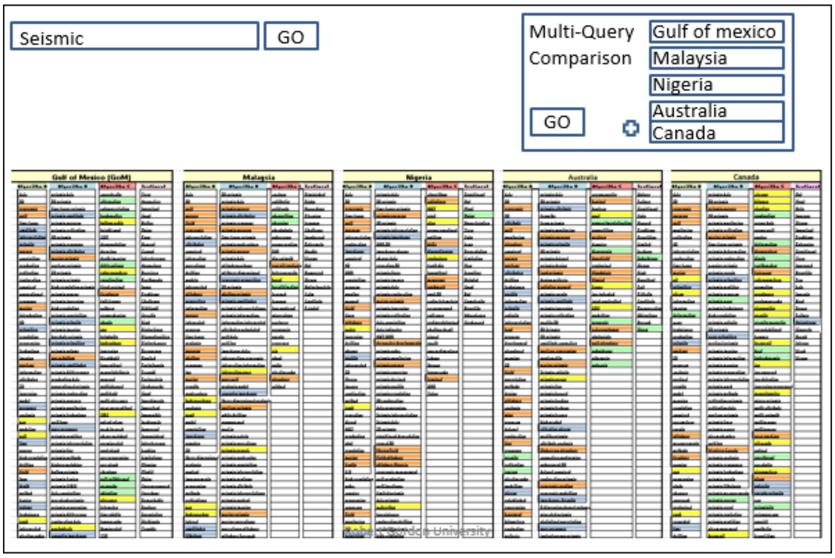
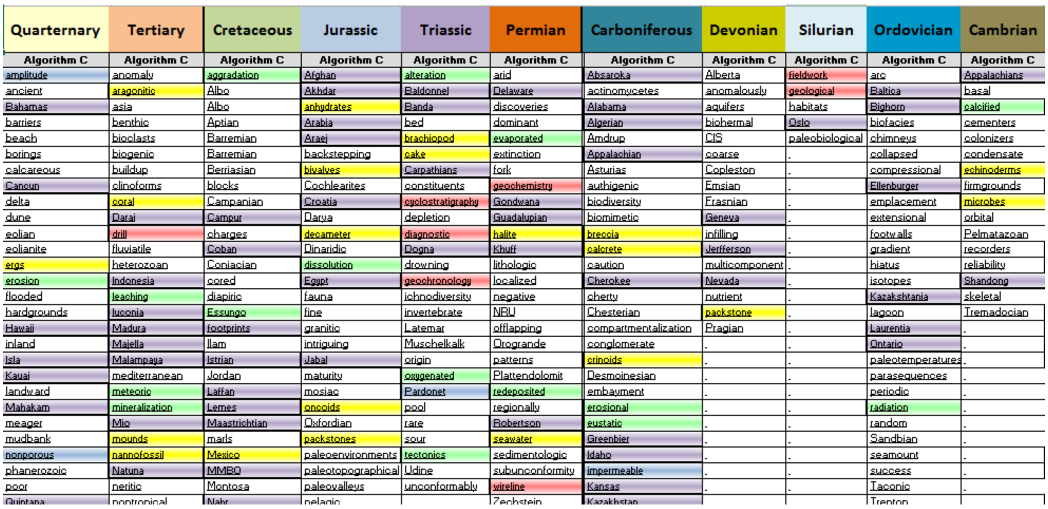


Figure 6 – Conceptual search user interface schematic used for organization #1. It is not intended for all the text to be readable, its purpose is to provide an overview of the concept. There were five secondary queries (e.g. Gulf of Mexico, Malaysia) each containing the lists created by the four algorithms. See Table 1 for an example of the terms.

The coloured visualization is inspired by DNA sequence displays. Each term was hyperlinked using Microsoft Excel to a corresponding URL which (when clicked or touched) would filter results and take the searcher to the document results online as Internet connections were available at both organizations. This was achieved using parameter parsing URL's for SPE's online library OnePetro (https://www.onepetro.org) and the Highwire search engine (http://home.highwire.org/) from Stanford University for GSL and AGI content. Participants were therefore able to identify an interesting association and click through to read the document(s) context that created the association.

The semi-interactive stimulus did not allow participants to enter their own search terms which is a limitation, but was deemed sufficient to stimulate needs for this research without the need to build a fully working prototype at this stage. The approach was keyword based so synonyms and hyponyms (Cleverley 2012) of the primary and secondary search terms within the text would not be identified and disambiguation of polysemic terms was not included. This was not necessary for the aims and hypotheses tested in this research study.



Quarternary	Tertiary	Cretaceous	Jurassic	Triassic	Permian	Carboniferous	Devonian	Silurian	Ordovician	Cambrian
Algorithm C	Algorithm C	Algorithm C	Algorithm C	Algorithm C	Algorithm C	Algorithm C	Algorithm C	Algorithm C	Algorithm C	Algorithm C
amplitude	anomaly	aggradation	Afghan	alteration	arid	Absaroka	Alberta	fieldwork	arc	Appalachians
ancient	aragonitic	Albo	Akhdar	Baldonnel	Delaware	actinomycetes	anomalously	geological	Baltica	basal
Bahamas	asia	Albo	anhydrates	Banda	discoveries	Alabama	aquifers	habitats	Bighorn	calcified
barriers	benthic	Aptian	Arabia	bed	dominant	Algerian	biohermal	Oslo	biofacies	cementers
beach	bioclasts	Barremian	Araej	brachiopod	evaporated	Amdrup	CIS	paleobiological	chimneys	colonizers
borings	biogenic	Barremian	backstepping	cake	extinction	Appalachian	coarse	.	collapsed	condensate
calcareous	buildup	Berriasian	bivalves	Carpathians	fork	Asturias	Copleston	.	compressional	echinoderms
Cancun	clinoforms	blocks	Cochlearites	constituents	geochemistry	authigenic	Emsian	.	Ellenburger	firmgrounds
delta	coral	Campanian	Croatia	cyclostratigraphy	Gondwana	biodiversity	Frasnian	.	emplacement	microbes
dune	Darai	Campur	Darya	depletion	Guadalupian	biomimetic	Geneva	.	extensional	orbital
eolian	drill	charges	decameter	diagnostic	halite	breccia	infilling	.	footwalls	Pelmatazoan
eolianite	fluviatile	Coban	Dinaridic	Dogna	Khuff	calcrete	Jerfferson	.	gradient	recorders
ergs	heterozoan	Coniacian	dissolution	drowning	lithologic	caution	multicomponent	.	hiatus	reliability
erosion	Indonesia	cored	Egypt	geochronology	localized	Cherokee	Nevada	.	isotopes	Shandong
flooded	leaching	diapiric	fauna	ichnodiversity	negative	cherty	nutrient	.	Kazakshtania	skeletal
hardgrounds	luconia	Essungo	fine	invertebrate	NRU	Chesterian	packstone	.	lagoon	Tremadocian
Hawaii	Madura	footprints	granitic	Latemar	offlapping	compartmentalization	Pragian	.	Laurentia	.
inland	Majella	Ilam	intriguing	Muschelkalk	Orogrande	conglomerate	.	.	Ontario	.
Isla	Malampaya	Istrian	Jabal	origin	patterns	crinoids	.	.	paleotemperatures	.
Kauai	mediterranean	Jordan	maturity	oxygenated	Plattendolomit	Desmoinesian	.	.	parasequences	.
landward	meteoric	Laffan	mosiac	Pardonet	redeposited	embayment	.	.	periodic	.
Mahakam	mineralization	Lemes	oncoids	pool	regionally	erosional	.	.	radiation	.
meager	Mio	Maastrichtian	Oxfordian	rare	Robertson	eustatic	.	.	random	.
mudbank	mounds	marls	packstones	sour	seawater	Greenbier	.	.	Sandbian	.
nonporous	nannofossil	Mexico	paleoenvironments	tectonics	sedimentologic	Idaho	.	.	seamount	.
phanerozoic	Natuna	MMBO	paleotopographical	Udine	subunconformity	impermeable	.	.	success	.
poor	neritic	Montosa	paleovalleys	unconformably	wireline	Kansas	.	.	Taconic	.
Quintana	nontropical	Nahr	pelagic		Zechstein	Kazakhstan			Trenton	

Figure 7 –Part of the visualization (poster) used for organization #2, Algorithm C only. The primary query (carbonate) and secondary queries (e.g. Tertiary, Cretaceous) with the top thirty ranked colour coded filters.

For qualitative data such as participant comments, a coding system was used to ensure participant anonymity. For example [O1G1_1] is the code identifying the first organization, first group and first participant. To avoid using acronyms through this article and for the sake of brevity, the phrase *discriminatory search term word co-occurrence* will be used to describe the set of comparison based techniques described in the previous section.

Results

In this section, the qualitative results from direct observation (audio and video-recording) and questionnaire comments along with quantitative data from closed questions and surveys in both organizations are reported for each of the participating organisations.

Organization #1 – Focus Group Qualitative Observations

The tension between information overload, whilst offering potential interesting associations was identified:
[O1G1_1] *"there is certainly scope for visualization of associations that I would not have had otherwise. The problem is how to reduce the information to just that bit that is most relevant. I feel least attracted to Algorithm D"* and *"Excitement was the first thought I had"* [O1G1_5].

Examples of serendipitous information discovery were identified, for example:
[O1G2_1]: "*The observation of carbonates in Malaysia is something that I was aware of, but did not immediately spring to mind when I think about seismic and Malaysia. Algorithm C made clear that I underestimated the importance of carbonates in Malaysia. It is immediately important for the exotic research that I am doing now, but it was relevant in my previous job as geophysical consultant."*

In the largest group in organization #1, an initial dialogue started up as the group gathered around the visual stimulus, relating to enterprise search in general. Although not specific to the visual stimulus, it surfaced a topical latent need regarding known item search, represented by the following interactions between three participants:

[O1G3_1]: *"I often think,… say something that you (looks at* [O1G3_2]*) don't want to hear. Depends on the data body behind it. If we had Google working properly on a full body of data we would be in better shape"*
[O1G3_2]: *"Why do you say that, that is an easy thing to say? What do you think Google will do for you to make it better?"*
[O1G3_1]: *"The key part is not Google, it is the full body of data, a good search engine on a full body of data"*
[Moderator]:*"You mean the Google experience? [nod of head from O1G3_1]"*

[O1G3_2]: *"Ah ok, so Google is a blinking word"*
[O1G3_1]: *"In the back of my mind, I think our problem is that there is a lot of data we don't have access to"*
[O1G3_2]: *"Now that I agree"*
[O1G3_3]: *"Really?"*
[O1G3_2]:*"I know for a fact that is true"*
[O1G3_1]:*"If I do a search on something I often don't find a document of which I know that exists"*
[O1G3_2]:*"yes, and we know why that is by the way"*
[O1G3_1]:*"Yah, permission is a big issue"*
[O1G3_2]:".*that is one, but another is the search is not indexing everything that is there"*

Participant [O1G3_2] is a Geophyscist by background but has had some data and information management responsibilities for a number of years, so may have a deeper understanding of this area than [O1G3_1]. This perhaps illustrates one of the challenges for exploratory search within an enterprise. If staff feel the *'bread and butter'* basics of *known item* search are not working effectively, it may be difficult to have a conversation about more advanced techniques. Participant [O1G3_6] spends most of their time teaching younger staff as part of the learning and development function. There was a general discussion about the fact a lot of information is in books which are not catalogued but may be stored in cupboards by senior staff or staff about to retire.

[O1G3_6]:" *I had a question this morning about how many elements are in the full gravity gradient tensor. They said they would Google it, but they could not find it. But I had a book, I knew where the book was. There are terms in our profession which are hard to find on Google. I Google everyday (every hour almost) to find things. Certain things in our profession though are really hard to find".*

Some of the participants clearly understood how word associations worked "*These words come out automatically*" [O1G3_4], whilst others struggled or were confused on how the associations were generated even after the introductory material "*This has expert's intelligence in it?*" [O1G3_5]. This may yield insights on differing levels of information literacy among the scientists. After this early exchange the focus went back to the visual stimulus:

An unanticipated topic was discussed. The geophysics discipline in organization #1 was developing a taxonomy. It appeared that nobody had thought (in addition to asking experts for terms) of using the data to automatically inform people about the terminology used in their information.

[O1G3_3]: "*an application of this we could be interested in is to help clean up. I could also see it could be extremely useful in the debate that is unravelling about the taxonomy, because taxonomy is difficult".*
[O1G3_1]:"*Yes this could help as a data driven taxonomy, very powerful".*

The uniqueness, non-obvious or unusual nature of words was of interest. During this time several participants touched the screen to reveal documents that contained the associations. Some discussion took place on this.

[O1G3_2]*:"This in itself, (I agree with [O1G3_6] initial search terms need to be more specific), but I know why you [looks at moderator] have chosen seismic as it fills the whole thing, but think it could help".*
[O1G3_6]*:"There is uniqueness.."*
[O1G3_2]*:"What do you mean?"*
[O1G3_6]:"*Well, uniqueness, like when I was looking for "wormy".. some of them attract my attention because they are very unique, most is not unique (e.g. seismic mapping), these are categories. I am looking for unique things that trigger my attention, this would be a starting point".*

A discussion on software tool functionality developed, with themes emerging such as interactivity, cascading and drill down to information. Some use cases for exploratory search were mentioned.

[O1G3_1]:*"I could envisage cascaded usage of this. So you first type in a term like seismic, it could then come up with seismic amplitude, you would click on that and it would do the same search again, maybe even triplet, in exploratory or discovery mode that you can zoom into something you find interesting. That is something I would probably do with this. This helps with big problem with Google (or that I have with Google), is choosing right selection of words to find something this tool could help you build up that selection of words".*
[O1G3_7]:*"Could you use OR or AND in the search terms [clicks and points at terms]?"*
[Moderator]: *"yes"*
[O1G3_1]*:"My feedback is build a real working version so we can play interactively, this is only semi-interactive"*

After one hour, all other participants had left the room, only [O1G3_6] and the moderator were left. The discussion then became more of a phenomenological style conversation. For example:

[O1G3_6]*:"I use Google as an exploratory tool. Something on the news hits me, I Google. I Google in the office as well, preparing for courses looking for lots of information. It is difficult to drill down into masses of information, this associative idea, may be something. Some terms not necessarily expert, difficult to get out of our data or out of Google, an example is synthetic, not even tied to our industry. If we can limit search to a domain, if that could be a boundary condition. Anything I type in, I get a lot back, but it does not help me in my search. In my mind I know exactly what the constraints are. I am searching constantly it is like I am doing nothing else."*
Moderator: "*What role does serendipity play in searching?"*
[O1G3_6]:*"I really like this, I like this associative stuff. I use this in class, I want to make people think. Associations are one way to get them to step out of their normal environment. It is like open up the box for me and I pick what does not fit with my brain, like one of those games*. It is also about context, like "inversion" is it geological inversion or geophysical inversion, many things are context sensitive".*

[*The comment "*like one of those games*" a type of "spot the difference" alludes to game playing elements in search.]

Organization #1 – Focus Group Quantitative Closed Questions

The hypothesis was that the approach could facilitate serendipity to a moderate/large extent. Within organization #1 75% of participants believed the approach could facilitate serendipity to a moderate/large extent. Using chi squared tests, this is statistically significant, ($p<0.05$) therefore the null hypothesis is rejected. The qualitative data supports this confirming the trustworthiness of the findings. Eleven of the twelve participants who thought the techniques useful, expressed a preference for algorithms C and D.

Organization #2 –Focus Group Qualitative Observations

Input from the focus group in organization #1 influenced the visual stimulus used in organization #2. For example, *"Frequency based algorithms (A and B) contained few surprises"* [O1G1_2]. More specific search terms were used and only the discriminatory Algorithm C presented. This provided the added benefit of allowing more secondary queries (twelve) to be used without making the display too large (wide). By the seventh focus group, few new themes were emerging indicating that a saturation point had likely been reached.

There was a remarkably different set of behaviours compared to organization #1. Participants seemed less interested in discussing the search issues within their own organization and focused immediately on the visual stimulus. Participant dialogue focused on what was unusual (to them):

[O2G4_1]: *"Interesting Majella pops up in the Teriary, but it has significance in Quaternary, so it's interesting it is most associated with Tertiary".*
[O2G4_2]: *"In here, it knows that in 15,000 articles, they have not seen any associations [with Quaternary]".*
[O2G4_1]:"*Anything else which is a bit odd?"*
[O2G4_3]:*"Cake is an interesting one in Triassic"*
[O2G4_4]:*"What does that refer to?"*
[O2G4_5]:*"Yes I have seen "intriguing" in the Jurassic. That is intriguing in itself" [group laughter]*
[O2G4_6]:*"Or just Egypt in the Jurassic."*
[O2G4_8]:"*What is interesting is Halite is there for the Permian, but technically it could occur for Tertiary, Triassic, Jurassic, every single one.."*
[O2G4_2]:"*So what is surprising is it hasn't…."*
[O2G4_3]:"*Silurian has not done very well" [long silence]*
[O2G4_2]:"*There is just nothing unique to it"*
[O2G4_5]:"*Habitats (for Silurian) is really weird" [confirmatory nods]*

During this discussion, participants used the tablet device to make searches on various associations. The discussion then moved from understanding what they were seeing, to conceptualizing what they would like to see:

[O2G4_2]:"*Say you are doing a search on biostratigraphy and you search for cretaceous would you want a list of these type of random things coming up to help you maybe search? [pause]*
[O2G4_7]:"*.. be interesting is finding Brachiopods in the Triassic, there is a lot of data, something to pick up on."*
[O2G4_8]: "*I really want to search on Brachiopods and see what comes up against all these columns, Tertiary, Cretaceous, Triassic, there will be loads of data"*
[O2G4_4]:"*Being able to split by region would be useful"*
[O2G4_2]:"*That would be better"*

[O2G4_5]:"*If you could run a species name through this that would be awesome for us"*
[O2G4_2]:"*Seriously?"*
[O2G4_3]*:"That would be pretty cool"*
[O2G4_2]*:"The key content is good way of showing what the most popular things at the moment being covered in different time zones and locations."* [Nodding of heads]

Most participants appeared to have a good understanding of what the algorithms were doing and the significance of the results. One conversation tackled the semantics of terminology being uncovered and again revisited the concept discussed in organization #1 of using the data to automatically help with dictionary/taxonomy development.

[O2G4_3]*:"Is there value in excluding certain words so for example with carboniferous you've Visean, but that's obvious as that's within the carboniferous so not going to give any information".*
[O2G4_2]:*"Yeah, maybe there should be a standard set of lookups to clean-up the display. We are currently creating a dictionary of all formation names and their aliases (also known as) using a tool like this may be a good way to narrow things down"*
[O2G4_4]: *"Can you link two works together like paleo spelt two different ways?"*
[Moderator]: *"It is possible to do that through a dictionary or through techniques like second order co-occurrence".*

Analogues were mentioned as a particular need that could be serviced through this type of approach.

[O2G4_2]:"*Is there any way we can use this to refine our search, or is it more useful ..to clean-up our own data?"*
[O2G4_4]:"*There is potentially some uses here for analogues. Getting something potentially useful by time intervals and geographical."*
[O2G4_5]:"*That could be quite useful"*
[O2G2_3]: *".. with analogues you don't know what terms to query on, because you don't know what they are"*
[O2G2_3]: *"It could be useful for finding analogues like finding Jurassic Rift Basins with Carbonate Reservoirs. You would not know where they occur geographically without prior knowledge. It would also be useful for finding how global events affect stratigraphy world-wide e.g. Jurassic Oceanic Anoxic Events (OAE)."*

Competitor Intelligence (CI) was another theme discussed as a possible use for the technique. The need here is to gain intelligence about the tactical and strategic directions of organizations. This can be hypothesis based (e.g. is company Y doing different things than company X in this geological basin?). Intelligence can also be data driven (e.g. these are the techniques being performed by company Y in geological basin Z) stimulating a hypothesis. Competitor intelligence information can be collected directly by a company and/or subscribed to through an information provider.

[O2G4_2]:"*We could type a company name [redacted name] into the search and look at the associations by geological age, so we could get a feel for where it is focusing its activities. We could also compare our own notes and research against the public domain to see if we were missing anything important like global events".*

Most groups asked several questions about the colours in the displays and how terms were categorized. Colouring was seen universally as a useful feature "*eye catching, spotting concepts of interest*" [O2G1_3], "*really helps to pick out*" [O2G4_8], "*visually much easier to correlate*" [O2G2_1].

Organization #2 – Focus Group Qualitative Survey Comments

No statistically significant differences were identified between the control and other groups, indicating that it was unlikely the moderator skewed the survey responses at discussion time. The following themes were identified through thematic mapping of the questionnaire comments using an approach based on grounded theory (Table 2). The comments were in response to the questions:

- To what extent do current search interfaces used in your organization facilitate serendipity?
- In your opinion to what extent can discriminatory search term word co-occurrence facilitate serendipity?
- If such tools/techniques existed in your company would you use them?
- To what extent does the colour coding affect the presentation of the visualization?
- Please provide suggestions for improvements and comments?

Characteristics were identified which would drive an intent to use and conversely, affect the intent to use such a tool. In addition, a value theme emerged that spanned both efficiency and creativity.

Table 2 – Thematic mapping of respondent comments in organization #2

Theme	***Description***
Business trigger (Task based)	At project start (project framing), for researching - background and literature searches. At the end of a project (After action review)
Business Trigger (Cognitive based)	When reaching a mind-block or getting stuck. Question forming, need to test a hypothesis generated during work. Need for more unusual material.
Intention to use (Value add)	Rapid way to see many associations, created linked data. Helps find analogues when the searcher does not know the search terms to use. Colours can help discriminate in an easy format to quickly evaluate. Unexpected patterns and associations, generate un-thought of topics and subjects.
Intention to use (Constraints)	A need to find a known item will not be helped by this approach. Perceived errors in the display (during colour coding, or semantic variants of same term). Amount of terms could overload and distract.
Value (Efficiency)	Faster access, time and effort saving.
Value (Creativity)	Help generate more insight, understand hidden connections, detect what is missing; lead to more investigation. Idea generation and prompting.

Organizational #2 – Focus Group Quantitative Survey

Question #1: To what extent do search interfaces within your organization facilitate serendipity? (Figure 8)

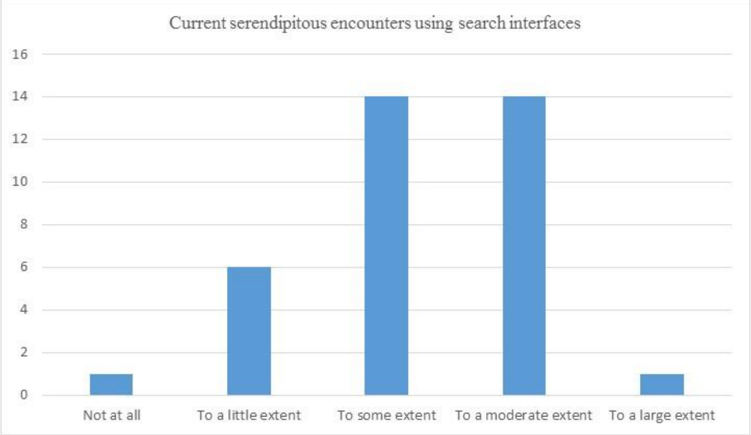


Figure 8 – Organization #2, question on current status, participant responses (sample=37)

Question #2 –To what extent could discriminatory search term word co-occurrence facilitate serendipity?

All respondents thought the discriminatory search term word co-occurrence techniques can help generate serendipitous experiences in some way (Figure 9). The hypothesis was that the approach could facilitate serendipity to a moderate/large extent. Within organization #2 73% of participants believed the approach could facilitate serendipity to a moderate/large extent (Figure 9). Using chi squared tests, this is statistically significant ($p<0.05$), therefore the null hypothesis is rejected. The qualitative data supports this, confirming the trustworthiness of the findings.

The contrast to feelings on the current state in their organization is striking with only 41% feeling current search interfaces have a moderate to large impact on their ability to find information serendipitously. Only one participant thought current approaches could help facilitate serendipity to a large extent, whilst sixteen participants thought the techniques used in the visual stimulus could. This is statistically significant, supporting the inference that the techniques could enhance current search interfaces in the enterprise. These results lend credibility to the notion that serendipitous encounters can be facilitated, i.e. they are more likely to occur in one search interface than another. It is not random.

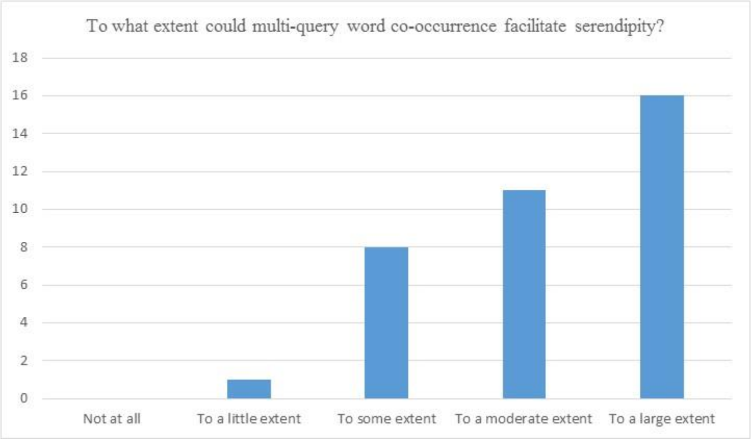


Figure 9 – Organization #2, question on potential, participant responses (sample=37)

The data in figure 8 and 9 is combined in Figure 10. The ordinal Likert scales have been assigned a numerical number. Although plotted on an axis (y) with equal distances between numbers, there is no assumption that intervals between categories are equal, the main purpose is a relative comparison. Number assignment is as follows: 1=not at all, 2=to a little extent, 3=to some extent, 4=to a moderate extent and 5=to a large extent. The extent that respondents thought current interfaces facilitate serendipity is displayed on the x-axis, from 'not at all' (left) to 'a large extent' (right).

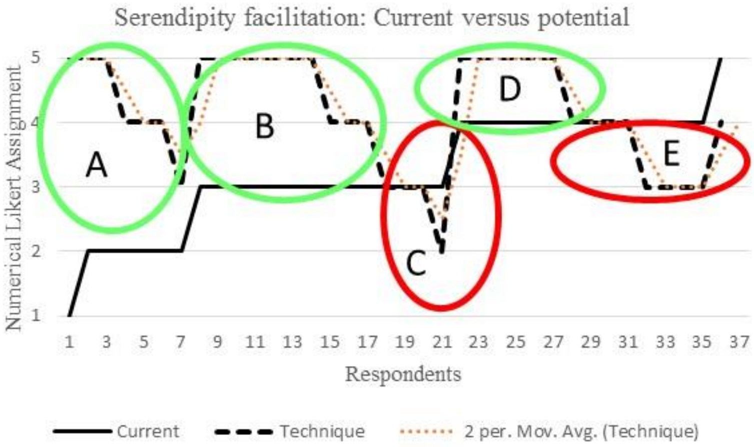


Figure 10 – Current ability of search interfaces in organization #2 to facilitate serendipity compared with discriminatory search term word co-occurrence techniques (dotted black line) based on the visual stimulus.

Five clusters of respondent beliefs are presented in table 3.

Table 3 – Response clusters for serendipity assessments

Theme	***Belief characteristics***
Group A	Low estimations of current techniques in the organization to facilitate serendipity (if at all), high expectations (large/moderate) for the discriminatory search term word co-occurrence technique.
Group B	Some estimations of current techniques in the organization to facilitate serendipity, high (large/moderate) expectations for the discriminatory search term word co-occurrence technique.
Group C	Some estimations of current techniques in the organization to facilitate serendipity, little expectations for the discriminatory search term word co-occurrence to facilitate serendipity.
Group D	Moderate estimations of current techniques in the organization to facilitate serendipity, high expectations (large) that discriminatory search term word co-occurrence could facilitate serendipity.
Group E	Moderate/Large estimations of current techniques in the organization to facilitate serendipity, some expectations for the discriminatory search term word co-occurrence to facilitate serendipity.

Where the dotted black line dips below the solid black line, the respondents (whilst still believing the discriminatory search term word co-occurrence can facilitate serendipity) take an *incrementalist* view towards the new technique. Where the dotted black line is significantly above the solid black line, the respondents have a more *transformative* view towards the new technique, with respect to facilitating serendipity.

Question #3 – If such tools/techniques existed in your organization would you use them? The survey responses are shown in figure 11. This question was a form of cross correlation, as one would expect responses to be similar to question #2, in that if a technique could help generate serendipitous opportunities, it could be inferred respondents would want to use it in most cases. The data supports the response made in question #2.

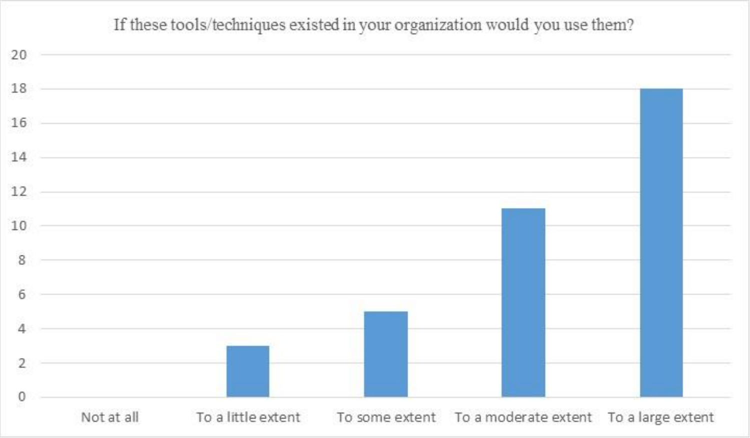


Figure 11 – Organization #2 question on use, participant responses (sample=37)

Question #4 – To what extent does the use of colour coding improve the display? This was in part a confirmatory question as some participants in organization #1 had liked the colouring aspect, however it was in part also exploratory as it may be seen as distracting by some participants. Figure 12 shows respondent answers.

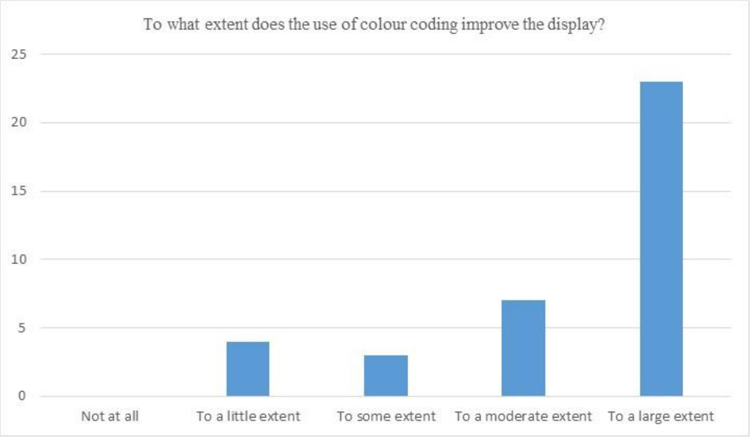


Figure 12 - Organization #2 question on colour, participant responses (sample=37)

The sixth question used a series of seven statements (McCay-Peet & Toms 2011) to allow respondents to describe their experience with the visual stimulus. The participants could tick all or none of the options to describe their experience. The statements are deemed to be a pre-condition to serendipitous "*aha moments*" effectively a measure of "*strategic insight*". No respondent ticked all seven statements, or left all blank. The average number of statements selected per respondent was between two and three, the mode was split equally (two and three) with the median number of statements selected being three. A radar chart showing responses is shown in figure 13. The responses seem to indicate that learning has taken place, even if it is not possible to explicitly document a serendipitous encounter in every case.

The most popular statement (68%) was "*I found things that surprised me*". Few participants (30%) indicated a need to follow up on those interesting association. This is lower than perhaps one would expect, based on the strength of responses around making new connections, obtaining unexpected insights and surprisingness. One explanation perhaps is that the terms in the stimulus were not chosen by the participants so they may be less immediately relevant to their work, than if they could use their own terms in the visualization.

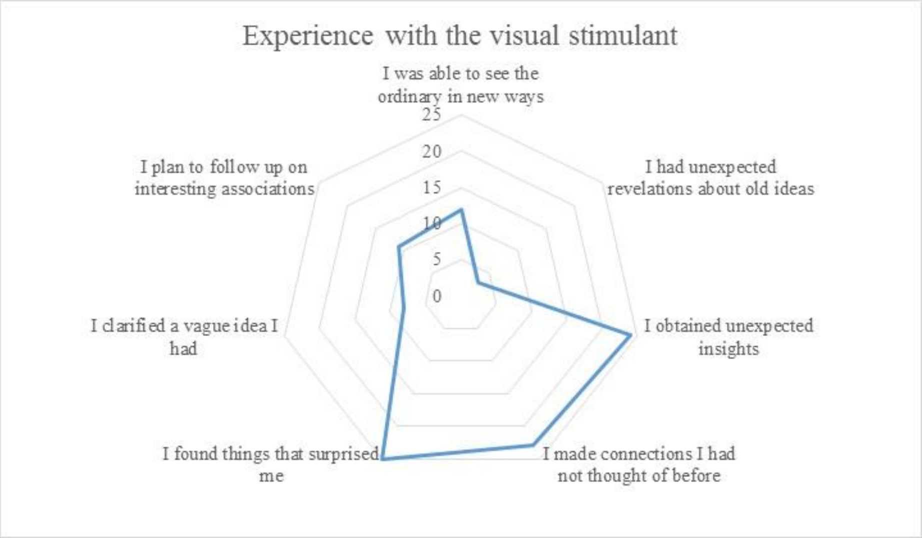


Figure 13 – Respondents experiences with the visual stimulus, sample=37

Discussion

In order for staff to adopt new approaches, there may need to be a compelling reason to break a comfortable searching style. With this in mind, respondents who indicated they felt the new approach could facilitate serendipity only to a little or to some extent were classed as ‘no use’.

Internal validity – organization #1

Of the twelve staff (75%) who indicated they would use the techniques if they existed, not all may use it just for exploratory search. It appears that taxonomy development is one area where participants saw value in some of the techniques outlined by the visual stimulus.

Some of the staff that felt the approach was of marginal value, seemed to have particular difficulty finding information that they knew existed within the organization. It could be that they have firm beliefs about what is needed to ‘*fix’* their particular issues and have little appetite to use more advanced techniques until these have been resolved. This could have important consequences for the relationship between the maturity of enterprise search in the organization and take up of new exploratory search techniques.

There appeared to be some participants who may fall into a *totalitarian-search belief*, focused on Google-like search “*Good results with Google struggle to see how this is a clear improvement*” [O2G1_4]. This belief may have formed as a result of dissatisfaction with enterprise search projects or initiatives within their own organization, combined with a strong tried and tested trust model for Google. This stance is characterized by people that use Google very frequently (home and in the office), frustrated at their own organization’s inability to deliver the same experience and want a single place for search believing Google (or equivalent) is the answer. They have little appetite for complimentary approaches to the classic search box, popular ranking and ten blue links and skeptical of things that look different.

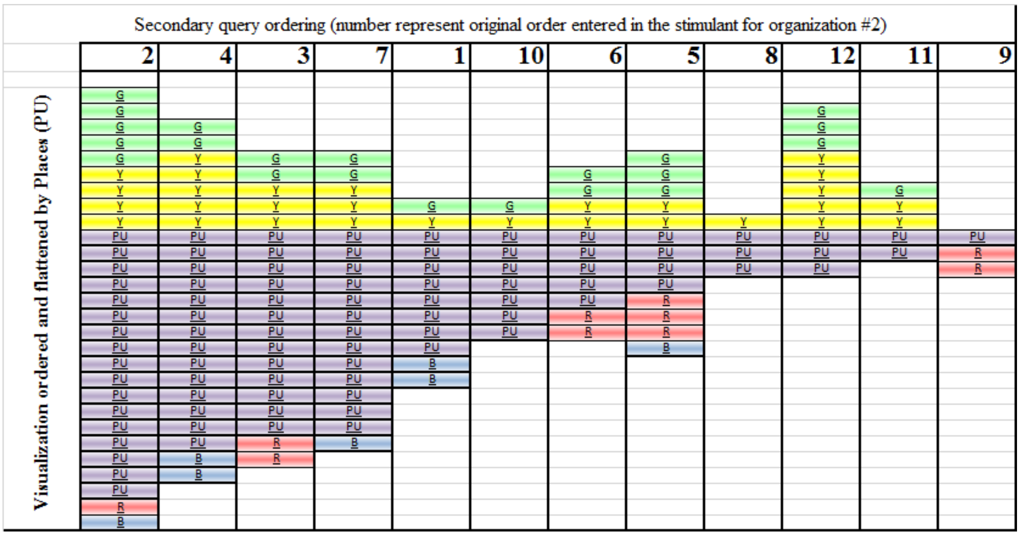
Some participants may fall into a *pluralist-search belief.* It is possible some of these beliefs have been influenced by the latter. This stance believes Google offers a great search experience, but does not solve all information search needs “*This would be a big improvement to my current search methods*“[O2G3_4]. They have an open mind, without necessarily having any evidence that there is a better way than the Google approach. They also use Google very frequently (home and in the office) but some may have a deeper understanding (than the *totalitarians*) of the complexities of information in an enterprise and the technological differences between Google as a technology and Google as an Internet search phenomenon (White 2012). Some may have a straightforward view that Google simply does not help them when they don’t know what search terms to use. The discourse between these two groups seemed to underpin some of the attitudes shown towards new exploratory search approaches. The almost exclusive use of “I” compared to “We” in the transcripts may indicate an absence of organizational collective beliefs in this area.

Participants within both organizations expressed positive views towards the *cited by* approach used by the Google Scholar type approaches working off peer reviewed literature. "*You look for a topic in Google scholar and you realize others are studying it from a completely different perspective"* [O1G1_1]. Smith (1964) coined the phrase *systematic serendipity* when reviewing the *Science Citation Index*. Smith understood that whilst in a keyword search the searcher may not be able to predict what would be retrieved, it would probably not be surprising. However, a system that allowed browsing of papers that cite other papers would probably lead to unexpected connections.

There were some desires to see such approaches applied to their own organizational content. A possible challenge to apply these techniques in the enterprise is the informal nature (with respect to referencing) of large amounts of enterprise document content. A *Google-scholar type approach* (as a concept) based on *cited by*, would not necessarily translate successfully into an organization as a means to discover connections and explore project documentation. Linked data (including usage data) may help to a certain extent, but other methods may be needed to create networks of associations to aid exploratory browsing. This lends credibility to some of the co-occurrence methods used in this research article.

Internal validity – organization #2

Some respondents had specific questions and preferences around presentational functional requirements for an interactive tool to deliver the techniques. These included the ability to sort the columns by word co-occurrence frequency and category (colour) both horizontally and vertically, so they could visualize certain patterns more easily. Figure 14 illustrates this need using data from figure 7. This is consistent with a view where participants were seeing value and forward thinking of the utility of such a tool and how they wanted it to work in practice.



Secondary query ordering (number represent original order entered in the stimulant for organization #2)

Visualization ordered and flattened by Places (PU)

2	4	3	7	1	10	6	5	8	12	11	9
G											
G									G		
G	G								G		
G	G								G		
G	Y	G	G				G		Y		
Y	Y	G	G			G	G		Y		
Y	Y	Y	Y			G	G		Y	G	
Y	Y	Y	Y	G	G	Y	Y		Y	Y	
Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU
PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	R
PU	PU	PU	PU	PU	PU	PU	PU	PU	PU		R
PU	PU	PU	PU	PU	PU	PU	PU				
PU	PU	PU	PU	PU	PU	PU	R				
PU	PU	PU	PU	PU	PU	R	R				
PU	PU	PU	PU	PU	PU	R	R				
PU	PU	PU	PU	PU			B				
PU	PU	PU	PU	B							
PU	PU	PU	PU	B							
PU	PU	PU	PU								
PU	PU	PU	PU								
PU	PU	PU	PU								
PU	PU	R	B								
PU	B	R									
PU	B										
PU											
R											
B											

Figure 14 – Example of potential functional requirements (juxtaposition) from the stimulus used in Figure 7

As well as exploratory search, some other uses were identified relating to dictionary creation and clean-up of the geological reference data in their own company. Semantics and taxonomy was also discussed, with many staff having experiences of the same concepts described or named slightly differently in literature, which may have hindered their search tasks in the past. In addition to discriminatory associations, there appeared to be a need to simply display associations by the secondary queries. Although Algorithms A and B were removed for organization #2, it is possible for very specific primary queries (e.g. Brachiopod) organized by other specific secondary queries, non-discriminatory algorithms may also generate unexpected patterns and perhaps facilitate serendipity. Colour coding of associations was found by 62% of respondents to improve the visualization to a large extent (81% thought it improved the display by a moderate to large extent). Even respondents of the *totalitarian search belief*, thought this element added value, "*the colour element to me is the clear improvement over Google"* [O2G3_6].

External validity organization #1 and #2

There was good agreement between organization #1 and organization #2, with 75% and 73% respectively, believing the techniques could facilitate serendipity to a moderate/large extent. The convergence coding matrix (Table 4) shows data combined from a variety of methods from both organizations. A reasonable degree of agreement is shown (ten out of nineteen themes in full, three in partial agreement), indicating possible transferability of some of the findings within certain contexts.

The only two areas of obvious disagreement between the data are whether an After Action Review (AAR) would be a valid trigger to use these techniques and whether sentiment based terms are useful in this context. The AAR process (Morrison & Meliza 1999) is a formal organizational learning process developed by the United States (US) military in the 1990's, looking back at project milestones or close-out for aspects which went well and not so well and why. The process is widely used in the oil and gas industry. Some participants thought the word co-occurrence techniques could be useful to look back at operational project documentation to identify possible learning points that people were unaware of. Other participants did not necessarily see this potential. This could be an area for further research.

Table 4 – Convergence Coding Matrix for contextual factors in organization #1 and #2, qualitative and quantitative

Contextual theme	***Theme meaning & prominence***			
	AG	***PA***	***S***	***DA***
Current methods in search interfaces within the organization could be improved to facilitate serendipitous discovery	●			
Discriminatory search term word co-occurrence (the techniques) in search interfaces could help to a moderate/large extent to facilitate serendipity	●			
The techniques are useful for taxonomy development	●			
Issues searching for known items within the organization			●	
Business triggers (Project framing, analogues, generate ideas)	●			
Business triggers (After Action Reviews (AAR))				●
Business triggers (Competitive Intelligence)			●	
Techniques has some intrinsic game playing element			●	
There are quite differing information literacy levels for scientists		●		
Display can be both overwhelming & distracting and at the same time generate vast amounts of useful options	●			
Need for an interactive prototype	●			
Use of colour enhances the visualization	●			
Need for *Google scholar type* cited by exploratory approach on internal company information.	●			
Techniques have capability to generate the unexpected/spark intrigue	●			
Sentiment based terms can generate the unexpected/spark intrigue				●
Techniques facilitate new insights and serendipity	●			
In respect to the techniques presented and their ability to facilitate serendipity, there are different belief clusters, with incrementalists and transformists.		●		
There are different stances towards search (totalitarians and pluralists) within the organization.		●		
Time savings			●	
Total	10	3	4	2

Where AG=Agreement, PA=Partial Agreement, S=Silence and DA=Dissonance

There were conflicting views on the usefulness of single word sentiment based terms. Many of the words did not seem interesting or unexpected enough for most respondents, others identified some (e.g. wormy). The use of abstracts in this research study might have hampered a realistic assessment of the potential value of this method. Using body text may have surfaced more intriguing sentiment based terms (e.g. cataclysmic, accident, too hot, value erosion, too low, lack of skills). It is possible multi-word descriptive terms (Cleverley & Burnett 2014) may be more useful to describe sentiment to a degree which attracts people's attention; an area for further research.

The nature of the underlying content will play an important role in the ability of a visualization to generate enough meaningful unexpected terms. Content from just a single narrow discipline will perhaps offer less potential for serendipitous encounters than an interdisciplinary collection. The results of a visualization on a corpus of 250,000 full body text reports within the domain of interest, is likely to be quite different to a visualization on only 5,000 report titles and abstracts. The more volume, the better the visualization is likely to be. Higher content volumes may present additional implementation challenges, in terms of the co-occurrence network size and noise reduction.

Although not explicitly tested, there was little evidence other than the occasional emotive phrase when asked how participants felt - *"my first thought was excitement*" that personality has an effect on interactions. Personality has been shown to affect search activity (Halder *et al.* 2010, Heinstrom 2002) and it may affect the desire to try new exploratory search techniques. Further research may be useful in this area.

Answering the Research Questions

Can comparison of search results by secondary contexts using word co-occurrence facilitate serendipity in the enterprise?

The data has shown that the discriminatory search term word co-occurrence techniques can facilitate serendipity to a moderate/large extent generally exceeding current capabilities within search tools used by the organizations today.

How does this happen?

The serendipity-by-comparison technique invites interaction where there is a difference between the 'pattern' in the stimulus and the 'pattern' in the mind of the observer. Associations deemed unexpected, attract attention. These can be unusual words (e.g. wormy), unusual contexts (e.g. Afghan associated with Carbonate and Jurassic) or unusual discriminatory associations (Halite just in the Permian).

The most frequent co-occurring terms were not deemed surprising by the participants. This is supported by existing research (Chuang *et al.* 2012, Olson 2007) that the most popular or most frequent categories, tags or facet values are not always what is needed by the searcher. Clearly, what is considered *curious*, *unexpected* or *serendipitous* by one person may not be by another, however certain algorithmic methods may be more suitable than others to facilitate the serendipity phenomena.

The technique and use of colours can also help pick out entity types (e.g. places), which allow discovery of analogs hitherto not known about.

The game playing nature of the 'spot the difference' technique may help engage searchers in problem solving, building on our natural desires for learning and achievement. It is possible the technique has some inherent gameful design without artificially including gamification concepts such as leaderboards or virtual rewards (Hamari *et al.* 2014). From observations across both organizations, the interactions with the visual stimuli are broken down into four cognitive based phases (Table 5) that bear a loose mapping to existing experiential learning theories (e.g., Kolb 1984). Hand face gestures have also been interpreted from observations and video (Mahmoud 2011).

Table 5 – The four cognitive phases (roughly in sequential order) observed in organizations #1 and #2

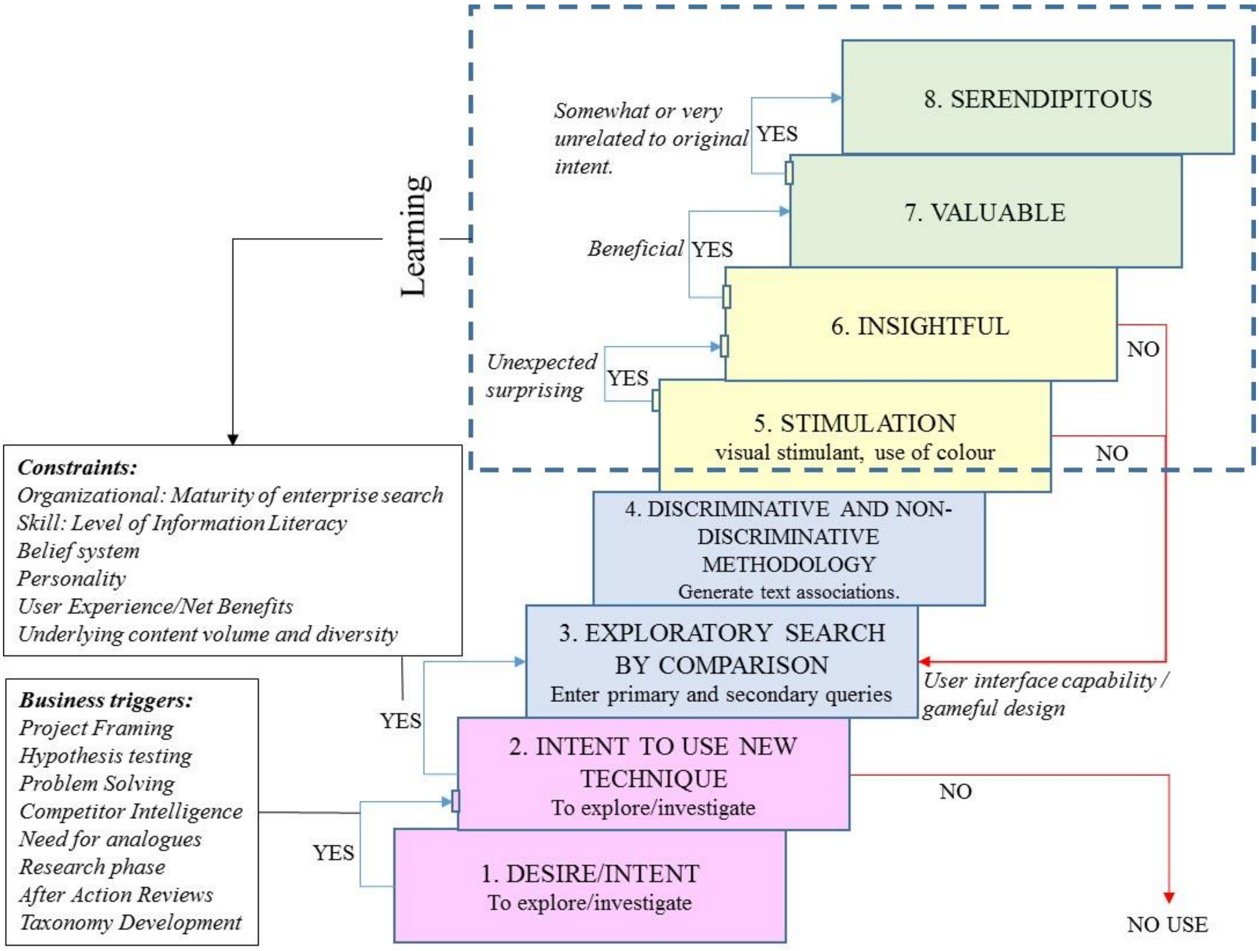
Theme	***Description***
Sensory overload	In the first thirty seconds some participants appeared somewhat *'overwhelmed*', perhaps *lost* - with a higher cognitive load than usual. It required orienteering for some participants.
Contemplation and comprehension	Participants scanned the display, thinking. Starting to *'get their eye in'*. Some more confident respondents started clicking on terms to view documents.
Clarification and information use	Respondents started to ask more detailed questions on techniques and specific terms in the display to clarify their understanding. Most respondents started to click on terms, drill down to documents; some were surprised.
Reflection and conceptualization	Participants moved away from the specific terms displayed in the visual stimulus and started to conceptualize what queries they would want to use for their own specific areas.

When would these techniques be of value to an organization?

Data gathered from the observations and questionnaires points to two main areas when these techniques are likely to add value for exploratory search. Firstly, task based triggers. These include at the beginning (literature review, start-up) or end of projects (reviews) or alternatively to support the competitive intelligence process. Secondly, cognitive based triggers when the searcher was unsure what search terms to use, is stuck (mind-block) or would like to test a hypothesis. The need for analogues could occur in either category. In addition the techniques were deemed useful to support one-off initiatives around taxonomy development and data clean-up.

Objective - Validating the theoretical model

The theoretical model appears broadly sound based on the data collected in this research. The model has been further refined to include a number of contextual factors (Figure 15).



8. SERENDIPITOUS
Somewhat or very unrelated to original intent.
YES
7. VALUABLE
Learning
Beneficial
YES
6. INSIGHTFUL
NO
Unexpected surprising
YES
5. STIMULATION
visual stimulant, use of colour
NO
Constraints:
Organizational: Maturity of enterprise search
Skill: Level of Information Literacy
Belief system
Personality
User Experience/Net Benefits
Underlying content volume and diversity
4. DISCRIMINATIVE AND NON-DISCRIMINATIVE METHODOLOGY
Generate text associations.
3. EXPLORATORY SEARCH BY COMPARISON
Enter primary and secondary queries
User interface capability / gameful design
Business triggers:
Project Framing
Hypothesis testing
Problem Solving
Competitor Intelligence
Need for analogues
Research phase
After Action Reviews
Taxonomy Development
YES
2. INTENT TO USE NEW TECHNIQUE
To explore/investigate
NO
YES
1. DESIRE/INTENT
To explore/investigate
NO USE

Figure 15 – Revised theoretical model to facilitate serendipity-by-comparison using discriminatory search term word co-occurrence

In the original model it was hypothesized that learning took place *because of* the serendipitous encounter. Subsequent to the research and the data collected, it is believed that learning can take place as soon as a term is deemed unexpected by the observer or simply the fact there is nothing unexpected is a learning by itself which may confirm an existing level of knowledge. If the intent to explore is highly related to the discovery, then value may be generated but it may not necessarily be a serendipitous encounter. Net benefits and satisfaction would be required to drive the intent to use such a system (DeLone & McLean 2003).

Conclusion

Discriminatory search term word co-occurrence techniques presented in this paper are capable of facilitating insightful and serendipitous learning opportunities and could be used to augment existing search methods used within organizations. The use of the techniques for taxonomy development and data clean-up were unexpected topics that emerged during the group discussions with the participants.

These techniques appear to offer a significant improvement over existing approaches used within the study organizations, providing further evidence that insightful and serendipitous encounters can be facilitated in the search user interface. Developing a *capability* that may lead to more opportunities for serendipitous encounters through the search interface is a wider construct than just technology. This will most likely include organizational culture, information literacy, maturity of metadata tagging and enterprise search in the organization, user satisfaction and net benefits.

Although the research study set out to investigate serendipity, it was discovered that the techniques may be useful to simply support exploratory based questions. If the intent is specific, then the results, even if surprising and valuable, may not necessarily be truly serendipitous. From an organizational learning perspective, if these approaches give rise to new insights and business value, whether they are classed as truly serendipitous or not, may be of secondary importance.

Although some examples of insightful and serendipitous encounters were documented, more empirical research within organizations with more interactive tools is required. The relative advantages and disadvantages of using these techniques in stand-alone approaches as opposed to integrating with existing precision focused search tools, is an area for further research.

The net benefits of new search approaches must offer clear benefits over existing methods if those methods are to be successfully disrupted and the *search-totalitarians* persuaded to adopt a more *pluralistic* stance. Those enterprises that develop enhanced capabilities in their enterprise search to facilitate insightful and serendipitous discovery which are subsequently *used* by enterprise staff, may gain significant business performance advantages over those that do not.

Acknowledgements

The authors would like to thank the Society of Petroleum Engineers (SPE), Geological Society of London (GSL) and American Geological Institute (AGI) for supplying their abstracts for this research. Appreciation is given to the two organizations who took part in the study, who remain anonymous for confidentiality reasons. A final thanks to Dr Laura Muir, Sylvie Davies, Professor Bill Kules and Dr Tony Russell-Rose for helping provide input into the article review.

References

Alexander, E., Kohlmann, J., Witmore, M., Valenza, R., Gleicher, M. (2014). Topic Model-Driven Visual Exploration of Text Corpora. IEEE VIS 2014, 9-14th November, Paris, France.

Allan, J., Croft, B., Moffat, A., Sanderson, M. (2012). Workshop report: Frontiers, Challenges, and Opportunities for Information Retrieval. ACM SIGIR Forum June 2012. 46(1), 9-11

Andre, P., Schraefel, M.C., Teevan, J., Dumais, S.T. (2009). Discovery Is Never by Chance: Designing for (Un)Serendipity. C&C'09, October 26-30th Berkeley, California, USA.

Appel-Meulenbroek, R. (2010). Knowledge sharing through co-presence: added value of facilities. Facilities, 28(3/4), 189-205

Bawden, D. (1986). Information-Systems and the Stimulation of Creativity. Journal of Information Science, 12(5), 203-216.

Beall, J. (2007). The value of alphabetically-sorted browse displays in information discovery. Library Collections, Acquisitions & Technical Services. 31, 184-194

Beresi, U.C., Kim, Y., Song, D., Ruthven, I. (2010). Why did you pick that? Visualizing relevance criteria in exploratory search. International Journal of Digital Libraries. 11, 59-74

Berniker, E. and McNabb, D.E. (2006). Dialectical Inquiry: A Structured Qualitative Research Method. The Qualitative Report. 11(4), 643-664

Bird, S., Klein, E., Loper, E. (2009). Natural Language Processing with Python. O'Reilly, USA.203-208

Brin, S. & Page, L. (1998). The Anatomy of a Large Scale Hypertextual Web Search Engine. Proceedings of the seventh international conference on the world wide web, 107-117.

Broder, A. (2002). A taxonomy of Web search. SIGIR Forum 36(2) 3-10

Bullinaria, J.A. and Levy, J.P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. Behaviour Research Methods. 39, 510-526.

Change, D., Dooley, L., Tuovinen, J.E. (2002). Gestalt theory in visual screen design. A new look at an old subject, in: Proceedings of the Seventh World Conference on Computers in Education: Australian Topics. (8), 5–12

Chapman, S., Desai, S., Hagedorn, K., Varnum, K., Mishra, S., Piacentine, J. (2013). Manually Classifying User Search Queries on an Academic Library Web Site. Journal of Web Librarianship 7(4), 401-421

Chuang, J., Manning, C.D., Heer, J. (2012). "Without the Clutter of Unimportant Words": Descriptive Keyphrases for Text Visualization. ACM Transactions on Computer-Human Transactions. 19(3)

Cleverley, P.H. (2012). Improving Enterprise Search in the Upstream Oil and Gas Industry Using a Non-Probabilistic Knowledge Representation. International Journal of Applied Information Systems (IJAIS).1(1), 25-32.

Cleverley, P.H. and Burnett, S. (2014). Retrieving Haystacks: A data driven information needs model for faceted search. Journal of Information Science

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. Psychological Reports, 80, 355-362.

Davenport, T.H., Prusak, L. (2000). Working Knowledge: How organizations manage what they know. Harvard Business School Press, Boston, USA. 60-61

DeLone, W. H. & McLean, E. R. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems, 19(4), 9-30.

Denrell, J., Fang, C., Winter, S.G. (2003). The economics of strategic opportunity. Strategic Management Journal, 24(10), 977-990

De Rond, M. & Morley, I. (2010). Serendipity: Fortune and the Prepared Mind. Cambridge University Press, UK.

Dessalles, J. (2009). Have you anything unexpected to say? The human propensity to communicate surprise and its role in the emergence of language. In Smith, A.D.M, Schouwstra, M., de Boer, B. and Smith, K. (Eds.) The evolution of language – Proceedings of the 8th International Conference (Utrecht), World Scientific, Singapore, 99-106

Fagan, J.C. (2010). Usability studies of faceted browsing: A literature review. Information Technology and Libraries. 29(2), 58-66

Farmer, T., Robinson, K., Elliot, S.J., Eyles, J. (2006). Developing and Implementing a Triangulation Protocol for Qualitative Health Research. 16, 377

Foster, A. & Ford, N. (2003) Serendipity and information seeking: an empirical study. Journal of Documentation. 59(3), 321-340

Friedman, B. (2010. Serendipity is an Explorationists best friend. American Association of Petroleum Geologists (AAPG) Online Article (http://archives.aapg.org/explorer/2010/04apr/mobilebay0410.cfm accessed June 2014).

Furnham, A. (2000). The brainstorming myth. Business Strategy Review. 11(4), 21-28

Gantz, J. & Reinsell, D. (2011). Extracting Value from Chaos (IDC). Report ID 1142

Ghiselin, D. (2010). Serendipity is alive and well at EagleFord. Hart's E&P Online Article (http://www.epmag.com/Exploration-Wildcats-Stepouts/Serendipity-alive-well-Eagle-Ford_73653 accessed June 2014).

Gwizdka, J. (2009). What a difference a tag cloud makes: effects of tasks and cognitive abilities on search results interface use. Information Research. 14(4)

Halder, S., Roy A., Chakraborty, P. (2010) The influence of personality traits on information seeking behaviour of students. Malaysian Journal of Library & Information Science; 15(1): pp. 41-53.

Halvey, M. & Keane, M.T. (2007). An assessment of tag presentation techniques. Proceedings of 16th International conference on World Wide Web. 1313-1314

Hamari, J., Koivisto, J., Sarsa, H. (2014) Does Gamification Work? – A Literature Review of Empirical Studies on gamification. In proceedings of the 47th Hawaii International Conference on System Science Jan 6th-9th 2014, Hawaii, USA.

Harris, Z. (1954). Distributional Structure. Word.10 (23), 146-162.

Hassan, A., White, R.W., Dumais, S.T., Wang, Y.M. (2014). Struggling or Exploring? Disambiguating Long Search Sessions. Proceedings of the 7th ACM international conference on Web search and data mining. 53-62

Haun, S. and Nurnberger, A. (2013). Supporting Exploratory search by User-Centered Interactive Data Mining. The 34th International ACM SIGIR conference on research and development in Information Retrieval, Beijing, China July 24th-28th 2011.

Hearst, M.A. and Stoica, E. (2009). NLP Support for Faceted Search Navigation in Scholarly Collections. Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP Suntec, Singapore 7th August 2009, 62-70

Heimerl, F., Lohmann, S., Lange, S. & Ertl, T. (2014). Word Cloud Explorer: Text Analytics based on Word Clouds. Proceedings of the 47th Hawaii International Conference on System Science (HICSS 2014), IEEE Computer Society. 1833-1842

Heinstrom J. (2002). Fast surfers, Broad scanners and Deep Divers – personality and information seeking behaviour. Doctoral dissertation Abo Akademi University. http://users.abo.fi/jheinstr/ (accessed June 2013).

Hoffart, J., Suchanek, F.M., Berberich, K., Lewis-Kelham, E., de Melo, G., Weikum, G. (2011). YAGO2: Exploring and querying world knowledge in time, space, context and many languages. Proceedings of the 20th international conference companion on World Wide Web (WWW), 229-232.

Hulme, T. (2012). Serendipity favours the connected. Online Article (http://www.wired.co.uk/magazine/archive/2012/09/ideas-bank/serendipity-favours-the-connected , accessed August 2014).

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

Kaizer, J. and Hodge, A. (2005). Aquabrowser Library: Search, Discover, Refine. Library Hi Tech News 2005. 10, 9-12.

Khalili, A., Auer, S., Ngomo, A.N. (2014). conTEXT – Lightweight Text Analytics using Linked Data. Extended Semantic Web Conference ESWC 2014.

Kolb, D. A. (1984). Experiential learning experience as the source of learning & development. Englewood Cliffs NJ Prentice Hall

Krestel, R., Demartini, G., Herder, E. (2011). Visual Interfaces for Stimulating Exploratory Search. JCDL'11 June 13-17 2011, Ottawa, Ontario, Canada.

Kules, B., Schneiderman, B. (2007). Users can change their web search tactics: Design guidelines for categorized overviews. Information Processing & Management. 44(2), 463-484

Kules, B., Capra, R., Banta, M., Sierra, T. (2009). What Do Exploratory Searchers Look at in a Faceted Search Interface? Proceedings of the 9th ACM/IEEE-CS joint conference on Digital Libraries. 313-322

Kules, B. & Capra, R. (2010). 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; 63(1), 114-138.

La Barre, K. (2011). Facet Analysis. Annual Review of Information Science and Technology. Information Today. 44(1), 243-284.

Leslie, I. (2012). In search of serendipity. The Economist: Intelligent Life. Online Article (http://moreintelligentlife.co.uk/content/ideas/ian-leslie/search-serendipity?page=full accessed September 2014).

Low, B. (2011) Usability and contemporary user experiences in digital libraries. CIGS Seminar, University of Edinburgh. Slide 17 http://www.slideshare.net/scottishlibraries/ux2-usability-and-contemporary-user-experience-in-digital-libraries

Lund, K., Burgess, C., Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. In. Cognitive Science Proceedings 1995.660-665. http://locutus.ucr.edu/Reprints.html (accessed October 2013).

Mahmoud, M., Robinson, P. (2011). Interpreting hand-over-face-gestures. Proceedings of the 4th international conference on Affective computing and intelligent interaction ACII'11. 2, 248-255

Makri, S., Blandford, A., Woods, M., Sharples, S., Maxwell, D. (2014). "Making my own luck": Serendipity Strategies and How to Support Them in Digital Information Environments. Journal of the Association for Information Science and Technology doi: 10.1002/asi.23200

Makri, S. & Blandford, A. (2012). Coming across information serendipitously – Part 1: Journal of Documentation 68(5), 68-705

Marchionini, G. (2006). Exploratory Search: From Finding to Understanding. Communications of the ACM. 49 (4), 41-46

Marshall, C., Rossman, G.B. (2006). Designing Qualitative Research. 4th ed., Sage, London. 114-115

Martin, K. & Quan-Haase, A. (2014). Designing the next big thing: Randomness versus Serendipity in DH tools. Digital Humanities 2014, 7-12th July, Lausanne, Switzerland.

McBirnie, A. (2008) Seeking serendipity: the paradox of control. Aslib Proceedings 60(6), 600-618

McCandless, D. (2012). Information in beautiful, 2nd ed., William Collins, London.

McCay-Peet, L. & Toms, E. (2011). The Serendipity Quotient. ASIST 2011, October 9-13, New Orleans, LA, USA

McCay-Peet, L. & Toms, E. (2011). Measuring the dimensions of serendipity in digital environments. Information Research 16(3)

Meyer, K.E., Skak, A.T. (2002). Networks, Serendipity and SME Entry into Eastern Europe. European Management Journal 20(2), 179-188

Michalski, R. (2014). The influence of color grouping on users' visual search behavior and preferences. Displays. 35, 176-195

Mindmetre. (2011). Mind the enterprise Search Gap, http://www.smartlogic.com/home/knowledge-zone/white-papers/1600-mindmetre-research-report-sponsored-by-smartlogic (accessed March 2013).

Minqing, H. & Liu, B. (2004). Mining and Summarizing Customer Reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 168-177

Morgan, D.L. (1997). Focus Groups as Qualitative Research: Planning and Research Design for Focus Groups. In Sage Research Methods, 32-46

Morrison, J.E., Meliza, L.L. (1999). Foundations of the After Action Review Process. U.S. Army Research Institute for the Behavioural and Social Sciences. Special Report #42.

Morville, P. and Callendar, J. (2010). Search Patterns. 1st, ed., O'Reilly, USA.

Nadeau, D., Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes. 30(1), 3-26.

Nitsche, M. and Nurnberger, A. (2013). Trailblazing information: An Exploratory Search User Interface. HIMI/HCII 2013, 230-239

Nolan, M. (2008). Exploring Exploratory Search. Information Architecture. Bulletin of the American Society for Information Science and Technology April/May 2008. 34(4), 38-41

Nunez, J.L., Lincoln, A., Rolnitzky, D. (2011). Manufactured Serendipity: Facilitating Accidental Innovation through a Web Application. UC Berkley School of Information May 5th 2011. Online Article (http://www.ischool.berkeley.edu/files/student_projects/smartsparq_paper_final_1.pdf , accessed May 2014).

O'Donnell, M. (2011). Visualizing Patterns in Text: Keynote talk at AESLA (Spanish Association of Applied Linguistics), University of Salamanca May 4th-6th. (http://www.uam.es/proyectosinv/treacle/Publications/AESLA-2011-ODONNELL.pdf , accessed September 2014).

Olson, T.A. (2007). Utility of a faceted catalog for scholarly research. Library Hi Tech. 25(4), 550-561.

Palmer, C.R., Pesenti, J., Valdes-Perez, R.E., Christel, M.G., Hauptmann, A.G., Ng, D., Wactlar, H.D. (2001). Demonstration of hierarchical document clustering of digital library retrieval results. Proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries, 451.

Raskin, R. (2011). National Aeronautical Space Administration (NASA) Semantic Web for Earth and Environmental Terminology (SWEET) Ontology. http://sweet.jpl.nasa.gov/ontology/ (Accessed February 2013).

Rasmus D.W. (2013). The Serendipity Economy. Harvard Business Review (HBR) Online Article. http://blogs.hbr.org/2013/08/how-it-professionals-can-embrace-the-serendipity/ (Accessed May 2014).

Reinanda, R., Odijk, D., de Rijke, M. (2013). Exploring Entity Associations over Time. Temporal, social and spatially aware information access workshop, August 1st 2013 Dublin, Ireland.

Rice, J. (1988). "Serendipity and holism: the beauty of OPACs". Library Journal, 113(3), 38-41.

Rivadeneira, A.W., Gruen, D.M., Muller, M.J., Millen, D.R. (2007). Getting our Head in the Clouds: Towards Evaluation Studies of Tag Clouds. Computer Human Interaction (CHI) '07 Proceedings Tags, Tagging and Notetaking April 28th May 3rd 2007, San Jose, California, USA.

Rose, D.E., Levison, D. (2004). Understanding User Goals in Web Search. WWW 2004, May 17–22, 2004, New York, New York, USA.

Rubin, V.L., Burkell, J., Quan-Haase, A. (2011). Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis. Information Research. 16(3)

Ruotsalo, T., Athukorala, K., Glowacka, D., Konyushkova, K., Oulasvirta, A., Kaipiainen, S., Kaski, S., Jacucci, G. (2013). ASIST 2013, November 1-6, Montreal, Quebec, Canada.

Russell-Rose, T., Lamantia, J. & Burrell, M. (2011). A Taxonomy of Enterprise Search. HCIR. 763 (Session 1), 15-18.

Russell-Rose, T. & Tate, T. (2013). Designing the Search Experience: The information architecture of discovery. Morgan Kaufmann, USA.

Sarrafzadeh, B., Vechtomova, O., Jokic, V. (2014). Exploring Knowledge Graphs for Exploratory Search. IIiX'14 August 26-29th 2014, Regensburg, Germany.

Scaiella, U., Ferragina, P., Marino, A., Ciaramita, M. (2012). Topical Clustering of Search Results. WSDM'12 February 8-12th, Seattle, Eashington, USA.

Shi, L., Wei, F., Liu, S., Tan, L., Lian, X., Zhou, M.X. (2010). Understanding Text Corpora with Multiple Facets.

Siefring, J., Roberton, I., Stewart, M., & Tessier, D. (2012). Problematic aspects of 'serendipity in information seeking.' Poster presented at the Serendipity, Chance and the Opportunistic Discovery of Information Research (SCORE) Workshop, Montreal.

Smith, J. F. (1964). Systematic serendipity, Chemical & Engineering News, 42(35), 55–56.

Soergel, D. (2009). Digital Libraries and Knowledge Organizations. In: Semantic Digital Libraries (ed. Kruk, S.R. & McDaniel, B.) Springer, 9-39.

Stenmark, D. (2008). Identifying clusters of user behaviour in Intranet Search Engine log files. Journal of the American Society for Information Science and Technology. 59(14), 2232-2243

Stasko, J., Gorg, C., Liu, Z. (2008). Jigsaw: Supportive investigative analysis through interactive visualization. Information visualization, 7. 118-132.

Strauss, A. & Corbin, J.A. (1998). Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory. 2nd Edition Sage Publications.

Streit, M., Schulz, H.J., Lex, A. (2012). Connecting the Dots: Linking relationships in data and beyond. IEEE Vis Week October 14th-19th Seattle, USA.

Sweeny, M. (2012).Delivering Successful Search within the Enterprise. British Computer Society (BCS) Information Retrieval Group. Online Article (http://irsg.bcs.org/informer/2012/01/delivering-successful-search-within-the-entepríse/ , accessed June 2013).

Taghavi, M., Patel A., Schmidt N. (2011). An analysis of web proxy logs with query distribution pattern approach for search engines. Computer Standards and Interfaces 34(11), 162-170

Teevan, J., Dumais, S.T., Zachary, G. (2008). Challenges for Supporting Faceted Search in Large, Heterogeneous Corpora like the Web. Proceedings of HCIR 2008.

Thompson, B. (2006). Serendipity casts a very wide net. British Broadcasting Corporation (BBC) Technology News. Online Article (http://news.bbc.co.uk/1/hi/technology/5018998.stm , accessed April 2014).

Thudt, A., Hinrichs, U., Carpendale, S. (2012). The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualizations. Computer Human Interaction (CHI) '12. May 5th-10th 2012, Austin, Texas, USA.

Toms, E.G and McCay-Peet, L. (2009). Chance Encounters in the Digital Library. ECDL'09 LNCS 5714, 192-202

University of Glasgow (UofG) (2013). Information Retrieval Group stop word list. Online http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words accessed November 2013

University of Nottingham (UofN) 2013. Survey unit: Focus Groups frequently asked questions (http://www.nottingham.ac.uk/survey-unit/focusgroupsFAQs.htm, accessed June 2014).

Vechtmova, O., Roberston, S., Jones, S. (2003). Query Expansion with Long Span Collocates. Journal of Information Retrieval. 6(2), 251-273.

Veling, A. and van der Weerd, P. (1999). Conceptual grouping in word co-occurrence networks. Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI). 2, 694-699.

Vicsek, L. (2010). Issues in the Analysis of Focus Groups: Generalisability, Quantifiability, Treatment of Context and Quotations. The Qualitative Report. 15(1), 122-141

Wertheimer, M. (1923). Laws of organization in perceptual forms. Translation published in Ellis, W. (1938). A source book of Gestalt psychology, 71-88 (http://psy.ed.asu.edu/~classics/Wertheimer/Forms/forms.htm , accessed March 2014).

White, M. (2012). Enterprise Search. O'Reilly, USA.19-20

Wilson, T.D. (1999). Models in information behaviour research. Journal of Documentation. 55*(3), 249-270.*

Wilson, M.L., Schraefel, M.C. (2008). Improving exploratory search interfaces: Adding value or information overload? In: Second Workshop on Human-Computer Interaction and Information Retrieval*, 23rd October 2008, Redmond, WA, USA, 81-84*

Wolfram, D., Wang, P, Zhang, J. (2009). Identifying Web Search Session Patterns Using Cluster Analysis: A Comparison of Three Search Environments. Journal of the American Society for Information Science and Technology. 60(5), 896-910

Yang, S.Q., Wagner, K. (2010). Evaluating and comparing discovery tools: how close are we towards the next generational catalog? Library Hi Tech. 28(4), 690-709

Yogev, S. (2014). Exploratory Search Interfaces: Blending Relevance, Diversity, Relationships and Categories. IUI'14 February 24-27 Haifa, Israel.

Zamir, O. and Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. Proceedings of the 8th international conference on world wide web, 1361-1374

Zhang, Y.C., O Seaghdha,, D., Quercia, D., Jambor, T. (2012). Auralist: Introducing serendipity into music recommendation. Proceedings of the 5th ACM international conference on Web search and data mining, 13-22.