International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 1– No.1, November 2012 – www.ijais.org

Improving Enterprise Search in the Upstream Oil and Gas Industry by Automatic Query Expansion using a Non-Probabilistic Knowledge Representation

Paul Hugh Cleverley
Flare Solutions Limited
12 Nelson Close, Wallingford
Oxfordshire UK OX10 0LG

ABSTRACT

Organizations face a vocabulary disconnect between the terminology people use in search and the inherent ambiguity of terminology in their information. The mismatch leads to critical information being missed. This paper discusses how Boolean keyword search, the most commonly used approach in Enterprise search, compares with automatic Query Expansion (QE) using a non-probabilistic Knowledge Representation (KR) created independently of the corpus.

The tests focused on the initial search results list. Optional recommendation or ‘what’s related’ options or facets were out of scope. Testing was performed on a globally created document library collection from one of the largest corporations in the world. QE recalled, on average, an additional 43% of relevant precise results in a single search, without a commensurate cost to information precision.

It is well known from set theory as more words are used in a keyword search, using an AND operator, fewer results are returned. However, it was observed as more words are used in a keyword only search, the relevant results returned, as a proportion of all relevant results in the corpus, decreases. This *narrow search paradox* means in general terms, when more search words are used in a query to help locate relevant information, as a proportion, more information of relevance is actually missed. This is caused by the compounding of words’ semantic fields and possible linguistic variants. It is believed this is the first time the effect has been modeled in this context, with wider significance in Information Retrieval (IR).

Keywords

Enterprise search, Web Digital Library, Query Expansion, Knowledge Representation, Information Retrieval, Semantic Ambiguity, Taxonomy, Ontology, Combinatorial Linguistic Explosion, Petroleum Exploration and Production (E&P)

1. BACKGROUND

Enterprise search commonly refers to a textual search engine deployed in a corporation that searches its unstructured or semi-structured information. This typically includes web pages, intranets, wikis, personal profiles, discussion networks and documents, in multiple geographies. These search engines have to cope with the continuing explosion of information. It is not unusual for a large multinational corporation to have over 50 million documents in its formal Electronic Document Management Systems (EDMS) alone. Supported by surveys 10 years apart (2001 and 2011 [1]), evidence suggests around half of all Enterprise search deployments do not meet users’ expectations in finding the information they need.

People have an amazing capacity to reason about the meaning of even tiny fragments of language, whether that is audio, visual or text. Search engines however, are not that intelligent.

It is unlikely that corporations will ever get to a situation where all relevant content is tagged systemically in a consistent way using common vocabularies. There will always be a vocabulary problem to overcome. In general, unstructured folksonomy style implicit and explicit tagging predominates. Having robust manual abstraction processes and strategies for publishing and tagging knowledge (metadata, annotations) is crucial. Using aliases from underlying systems to feed the search is crucial. Having retention policies to delete obsolete information is crucial. But these elements do not provide a total solution in themselves.

1.1 Human Interaction

Users’ search behaviour can be influenced by their experiences of internet search engines like Google. This expectation is carried into the workplace by staff. It often causes frustration, uncertainty and anxiety when they cannot find information. Staff perception is that they always find information using Google and there is an expectation that searching corporate information assets should be just as easy.

The users’ satisfaction of search quality goes beyond just a fast response from a large index with well ranked results. Presentation of results, summarization, notifications and refiners or facet navigation will play their part. The challenges of Enterprise Search are well documented by Hawking [2].

In *Google like* internet searches, as long as what the user is looking for is on the first page, it does not matter if masses of irrelevant results are recalled on later pages. Search in a corporation can have other requirements. Staff can require exact listings/reports of results that match queries. The provenance, currency and quality of information can be important to find. Searches are used to support information gathering, analogue identification, operations, asset acquisition & divestment and legal processes. Generalizing, search results need to be more precise with more attention to all the recalled results in a corporation, than on the internet.

There is also high variability in outcome, based on the background and cognitive skills of users. One user may type a three term query, find a number of internally relevant published reports and believe that they have found everything. A more seasoned individual may make more searches, using lexical and semantic variations, to tease out additional relevant reports. Different users may even interpret the same question differently, so retrieve different documents.

Age also plays a role. Younger people (<30 years), as opposed to those with 10-20+ years more experience, are more likely to perceive source selection and formulation of a search query are the most significant information seeking activities. Their biggest problem being search output is not exhaustive enough (Chowdhury et al [3]). This may be because experienced people, used to searching for information *before-Google*, have better insights into the limitations of search technology.

1.2 Information Overload

There is a trade-off between information recall (completeness) and precision (accuracy). The point at which retrieving additional (relevant) results degrades information precision is called the tipping point. This can be measured using the F-measure, the harmonic mean of recall and precision.

Certain types of statistical techniques can be used to retrieve similar documents that may not necessarily match all the words in the initial search query. Anecdotal evidence from corporations where these techniques have been used in the initial search tends to be one of over recall and poor precision. User expectations are not fulfilled. It is not uncommon for corporations to switch off or marginalize these types of statistical techniques in these engines, resulting in the Boolean keyword based methods observed in mass use today. There is sometimes confusion over whether these types of statistical approaches should be used as optional recommender or suggest mechanisms (for related documents), or embedded as part of the default initial search results list, or both.

1.3 Semantic Web and Domain Ontologies

Initiatives exist to make sense of the billions of web pages on the internet. The Semantic Web is a collaborative movement led by the International Standards Body, the World Wide Web Consortium (WC3). One of its aims is to make web pages more understandable for computers so they can carry out more complex or intelligent actions like search. Many ontologies (defined as a conceptualization of a specification, Gruber [4]), have been manually developed in certain industry sectors, some in collaboration with WC3. In the oil and gas sector, to speed up projects, the ISO 15926 Ontology was developed, focusing on technical data definitions and digital interoperability for the handover in capital intensive projects from the contractor to the operator of oil and gas production facilities. Norwegian oil and gas companies (e.g. Statoil) played a pivotal role. Recent initiatives are attempting to broaden the scope to chemical and process industries. Despite the business issues, the practical use of semantics to enhance Enterprise search has received little attention.

1.4 Oil and Gas Thesauri and Taxonomies

Numerous thesauri or taxonomies exist in the upstream oil and gas industry. For global geographical entities for example, I.H.S. Energy license hierarchical lists for wells, fields, licenses and basins etc. The US Geological survey has publically available sets. There are many more.

For non-geographical keywords, the publically available Schlumberger Oilfield Glossary has 5,000 terms. The Government of Western Australia's Geoscience Thesaurus (GeMPet) has 10,000 terms. The University of Tulsa has 13,000 non-geographical keywords which it licenses to search its own papers. DataFacet Inc licenses its oil and gas taxonomy of around 1,500 terms.

In 2002, an initiative from Shell and Flare Solutions released several thousand Document Types, keywords and related catalogue standards into the industry, under the custodianship of Energistics (the Energy Standards Organization). This publically available information, termed *EPICAT*, could be used to describe any piece of information. Non-geographical terms from this set form the core of a vastly expanded KR model of over 100,000 terms, which Flare Solutions commercially license. The Flare model was used for QE in this paper to simulate the effect of applying a large KR when searching a global document library.

1.5 Geospatial

Building on the generic *EPICAT* work, in 2009 work led by Chevron began focusing on creating a generic ISO standard to describe all information, the Energy Industry Profile (EIP) of ISO 19115-1 [5]. Initial focus is on geospatial information. This provides a framework for discovery, retrieval and interoperability of spatial layers. The ISO standard simply lists 19 classification topics for layers (e.g. *farming*, *health*, *oceans*, *intelligence/military*), although the Infrastructure for Spatial Information in Europe (INSPIRE) has 34 topics (e.g. *Buildings*, *Geology*, *Transport Networks*).

These high level topics aside, no other semantic keywords exists. The European Environment Information and Observation Network (EIONET) provide a thematic public thesaurus of hierarchical keywords (GEMET). This consists of a few thousand general keywords, with around 400 keywords covering Energy.

2. OBJECTIVES

It is a given that using certain linguistic variants not currently used in a search, will improve recall and keep precision. This study does not seek to prove that expectation or to investigate the merits of the KR used. It focuses on the *magnitude* of the additional relevant recall that can be typically discovered using a KR in QE in a large multinational corporation.

3. RELATED WORK

Query expansion (QE) (or augmentation) is the process of enriching an initial user query to improve IR. The ambiguity of language is the underlying reason why QE is applied. A key aspect within QE is reasoning, the process of using existing knowledge to draw conclusions. QE can be achieved by a variety of techniques, some are discussed here. In corpus dependent models, discriminant terms in the 'top ranked' search results (either chosen by the user (feedback) from the results, or automatically by the system) are used statistically for QE. The Latent Semantic Analysis (LSA) technique is commonly used to determine relatedness. On the basis that documents containing distributions of infrequent words, which are also present in the top ranked documents for a query, must also be related. Ogilvie and Callan [6] discuss the difficulties in achieving this in distributed (federated) environments, merging ranked results lists. In corpus independent models, external sources can be used (e.g. Wikipedia), termed Well-known Collection Enrichment (CE), Peng et al [7], using similar statistical techniques. In Topic modeling, topics are represented as probability distributions over the vocabulary. Yi and Allan [8] noted topics discovered from the whole corpus are too coarse for QE however topics from feedback documents appear to have potential.

The above methods are purely statistical, so applicable to any type of information. If QE is to be used in a highly technical domain, pre-defined Knowledge Representations (KR) can provide value if applied automatically. This could take the form of a vocabulary, thesauri, taxonomy or ontology. These can be created manually or semi-automatically, the emphasis is on quality. Vorhees [9] indicated that automatically using a thesaurus such as WordNet did little to improve search effectiveness. This is probably more related to the relationships in WordNet and the content it was being applied to. Kristensen [10] found recall was doubled when a thesaurus was used for QE with a 10% decrease to precision.

The Biomedical industry has numerous published papers where domain KRs, like ontologies have been used automatically for QE in search (Segura et al [11], Bhogal [12]). It is believed there is only one such paper on QE in the Oil and Gas Industry (Solskinnsbakk & Gulla [13]). Their experimental findings indicated search recall could be increased, but at significant cost to precision. One of the challenges they faced was using an ISO Ontology, not aligned with everyday language used in oil and gas documents.

Almost any concept is related to any concept in some way, so the relative importance of relationships between concepts is commonly represented as a relative weight. In an *"is-a"* taxonomy, semantic similarity decreases as you traverse the hierarchy away from the given concept. Resnik [14] proposes an information content based approach to calculate similarity, as opposed to how many edges a concept is away.

Algorithms such as Spreading Activation (Quillian [15], Collins and Loftus [16], Crestani [17]) biologically inspired by the cognitive processes of the human brain, can activate and combine these various weights. Probabilities can be derived for how semantically related concepts are, to a given activation query or input. Recurrent semantic networks (a weighted graph) are the backbone of this approach. Recent work from Wojtinnek and Pulman [18] on determining semantic relatedness and Liu et al [19] on extension of an existing domain ontology using these techniques could have many applications to improve Enterprise search.

4. BUSINESS PROBLEM

Searches made by a user can miss relevant information due to semantic mismatches. Some of the common causes for semantic ambiguity are given in Table 1. Users and support staff often try numerous variant queries to mitigate this issue. These practices are idiosyncratic, delivering incomplete results and wasting staff time.

Table 1 – common causes of semantic mismatches

Name	Example
Acronym	If a user makes a search on 'Light emitting diodes' they may miss a report on LED's.
Synonym (+Pseudo)	If a user makes a query on 'car', they may miss a report on 'automobiles'. These can be thought of as contextually interchangeable.
Hypernym / Hyponym	If a user makes a keyword query on 'Fruit', they may miss a report on 'Apples'. This is related to the level of language used, if searching using more general words.
Metonym	If a user searches on 'Oil Supermajor' they may miss reports on 'Big Oil' (where a figure of speech concept is not called by its own name but is culturally something intimately associated with the concept).
Meronym / Holonym	A user can make a search on 'car' but miss a report not mentioning car, but containing numerous (*part-of*) relationships such as steering wheel, chassis, engine, axel.

The problem is acute where only limited text is available to search (titles, descriptions, keywords). This includes physical library and third party information (copyright restrictions). This also applies to published electronic documents, published in this context, meaning all staff can see the existence of the item in search but the actual content or body text is not necessarily open for everyone to see by default (i.e. once found, some staff may need to request access). In these cases, it is only possible to search limited metadata (not the entire text). Even if all the body text of electronic documents can be searched, the title, description and keywords/tags provide the key to good ranking. Semantic mismatches in these elements are critical otherwise relevant content will be effectively *lost in place*. The business impact of missing information through these mismatches can lead to sub-optimal decision making for projects involving millions or even billions of dollars. It is not uncommon for analyses and decisions to be made, only to later find information that may have significantly changed conclusions or decisions. In many cases it will probably never be known what was missed - *you don't know what you don't know*.

5. METHOD

This paper focuses on comparing keyword and domain KR based QE, where precision is not significantly degraded. So for simplicity, the part of the KR chosen contained generally monosemic terms and non-conditional relationships. Probabilistic techniques were not therefore needed, as all weights were either 1 or 0. The amount of information missed by keyword searching (found by QE techniques using these parts of the KR) can therefore be considered a minimum. To use a cliché, the *low hanging fruit* were targeted.

5.1 Knowledge Representation Creation

Taxonomies of concepts were manually created using oil and gas domain knowledge. Each concept was modeled as a deep hierarchy (taxonomy) with synonyms and related to other concepts using pre-defined relationships (e.g. Topics were linked to certain Disciplines). Information Product Types (PT) refer to information items (i.e. data/document types), produced by routine *factory like* physical or knowledge processes. PTs have their own pre-defined relations (Fig 1).

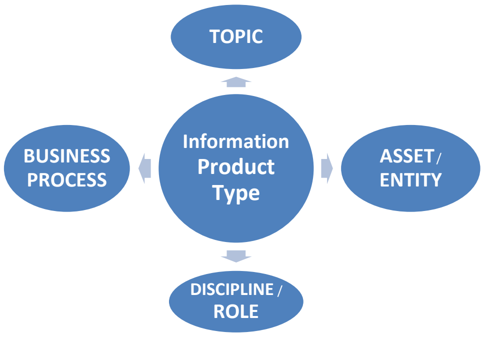

Figure 1 – PT Part of Knowledge Representation Model

Topics refer to types of materials, properties, techniques, features, problems, equipment etc. In Ontological nomenclature these are classes and sub-classes. Named Entities (NE) types (Assets) are instances (individuals) of certain types of topics (e.g. actual oil and gas wells, fields, technologies, organizations, etc). The associations (connections) made between concepts is the *semantic memory*. In probabilistic networks, the strength of the individual connection is represented by a Probability (P) or weight. The associations used in the study were definitional, so all had the same weight, effectively P=1.

Oil and gas corporations have a nebular scope of interest. This requires the conceptual importing of upper ontological elements into a domain, which is a quasi-combination of a number of part-domains. Scope includes *Geoscience, Engineering (various), Operations, Commercial, Finance, HSE, R&D, HR, Legal, Project and Business Mgmt., Geomatics, Metocean, Data/Info/Knowledge Mgmt.*, etc.

The manually created concepts were embellished with the results of automatic approaches using parsers, co-occurrence and subsumption techniques. Several thousand public domain oil and gas papers were used as the input. The automated approaches took the manually created concepts as a seed and used lexical stemming and proximity algorithms to look at associated words commonly found in text adjacent to the seed concepts. The resulting associations were stored in multi-dimensional vectors. The most monosemic prevalent concepts and relationships to the seed were manually analyzed and fed back into the model and iterated further. This, in effect, added new concepts and/or associations. Section 8 discusses how future activities could automatically mesh more probabilistic relationships onto the existing definitional framework.

It was observed that the majority of relationships automatically learnt through content were a measure of semantic relatedness, as opposed to semantic similarity. This perhaps illustrates the point made by Velardi et al [20], that it is virtually impossible to automatically recreate highly technical taxonomies from document content.

The result was 25,000 concepts (100,000 terms) and half a million pre-defined relationships linking concepts together. Domain oversight, although time consuming, provided quality and meaning, mitigating the noise and coverage issues seen in clustered concept hierarchies automatically created from text (Woon & Madnick [21]). Fig 2 shows the number of words used in Topics and PT concepts, having a median of two words and three words respectively. The latter's average is closer to four words.

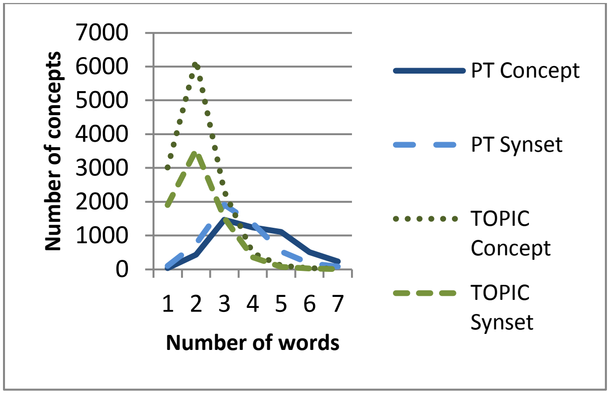

Figure 2 – Number of words in concepts within the KR

Broadly speaking, where the concept hierarchy is based on semantic similarity, it can be described as taxonomic. Where the concept hierarchy is based on semantic relatedness, it is more ontological in nature. Semantic similarity is a form of semantic relatedness. For example, the Discipline *Health, Safety & Environment (HSE)* and Topic *Geohazard* are more semantically related than the two Disciplines *HSE* and *Finance*. However, *HSE* and *Finance* are more semantically similar (than *HSE* and *Geohazard*), because they have an '*is a*' relationship to Discipline. Relatedness between concepts shares a different type of meaning. Combining both of these in the KR model avoided a biasing to just semantic similarity which often happens when only taxonomies are used for QE.

5.2 Content Indexing

The Enterprise search engine in production use within the corporation was used to index the corporate document library of 170,000 items (the corpus). The documents were in English but produced by people in many different countries and roles. The reports consisted of titles, descriptions and keywords. Many of the latter were added by users, so are a form of folksonomy. This resource represents a significant investment of time and money for the corporation stretching back decades. Several thousand third party journals subscribed to by the corporation were also included (these could have been any type of external content e.g. Patent information).

Two indexes were created, one which only contained the original text (for keyword search) and one which contained the semantically added QE terms. Some logic was added to the QE results lists so they only showed the unique difference to keyword search (Δ), for the given query.

5.3 Inductive and Deductive Inference

A basic set of IF THEN inference rules were applied within the KR itself, to create artificial synonyms and examples.

As the document text was indexed, it was parsed against the KR and a second set of inference applied. So corpus QE was effectively done *up front* for performance reasons, using a very simple (order independent) matching algorithm. The whole sentence was parsed, so compound concepts were identified where they matched those in the KR.

5.4 Search Configuration

Free text was used to query the results of indexing and inference. Stemming effects were normalized, so additional results found through QE would not be related to simple morpheme lexical variants (e.g. flood v floods v flooding).

Without normalization, it would be possible to show improvements in search recall using QE techniques, but many of the additional results could be produced by lexical, not semantic means. Lexical variants are often found by users through the use of wildcards (typically * or %) when searching. Random spelling mistakes in the query or corpus were not addressed.

Spaces between search terms were treated as a Boolean 'AND' operator, which is standard practice in most search engines. No query terms were automatically dropped (even if it delivered zero results) so strict logic was applied.

Search ranking included proximity. Where an exact match on search terms was found in a document (in the original or semantically added text) it was ranked higher than a match where the given terms were not in proximity. This had the general effect of making the first or 'top' set of results generally quite precise.

Time constraints did not allow changing of the actual search query made, by the type of concept recognized. For example, for a query made which included a recognized concept "Well Proposal", it may not be necessary to search explicitly for its constituent words "Well" AND "Proposal". See Section 6.3.

5.5 Number of Test Queries

The number of queries used to test the various methods was based on research by the Text Retrieval Conference (TREC) [22] that originated from the Defense Advanced Research Project Agency (DARPA). When comparing search engines, their work indicated that using 25 queries resulted in a 13% margin of error. This was enough to change the relative position of one search engine's performance against another. Doubling the number to 50 test queries reduces the margin of error to 4% so 50 queries were chosen for this study.

5.6 Number of Terms in Test Queries

The number of terms used in test queries was based on two lines of evidence. According to Taghavi et al [23] in a 2011 study of internet proxy logs, 70% of all Internet search queries were found to be three terms or less. This is up from around two terms in 1996. Data from digital libraries (Nanoscience) (Shiri and Chambers [24]) indicates 89% of all queries were three terms or less. Sometimes a source is pre-selected by a user prior to a search (e.g. a certain sub-collection of a library, or EDMS instance/site), so searches have additional context.

5.7 Types of Test Query

A range of domain discipline staff in the corporation provided and verified the actual search queries used. The types of search queries were split into two main types. Firstly, information work product type queries, which typified reports or documents that get repeatedly produced from business processes. These typically ended up with a term at the end such as *data*, *proposal*, *plan*, *programme*, *analysis*, *map*, *test*, etc. Most people will recognize this nomenclature as document types and data types. Secondly, topic based query types such as *processes*, *features*, *models*, *techniques*, *properties*, *materials*, *equipment*, *problems*, etc. A real world (instance) such as specific geography or asset (entity) was often combined in the search query. Natural language queries were not tested.

5.8 Measuring Precision

Precision (relevance) is subjective, defined as how well a retrieved information item or listing meets the need of the information consumer. Precision or relevance does not indicate whether all relevant information has been retrieved. In theory, how relevant a result is, depends on how semantically related it is to the given query term(s). Staff with domain knowledge carefully analyzed the results set to assess the relevance of each result. Precision is often given *in terms of k*. Where k is the number of results analyzed, k=10 would be precision of the first page only, where there are 10 results to a page. Query precision was measured at k=10, and for the entire result set as a whole for both keyword and QE results.

6. RESULTS

6.1 Information Precision

The keyword search results had a precision overall of 95% and at k=10 close to 100%. QE results were less precise, with overall precision of 81%, precision at k=10 of 90%.

6.2 Relevant Information Recall

For each query, a percentage was calculated for the additional relevant results found by QE (missed by keyword search). Queries were treated equally. A keyword search finding 20 results where QE found 80 results (80% additional) was treated equally to a keyword search finding two results where QE found 8 results (80% additional). The average for 50 queries is shown in Table 2. For example, a keyword search on *Enhanced Oil Recovery* (EOR) for a given geography found 36 results. Using QE another 145 unique relevant results were returned, including documents on *Surfactant flooding* that did not mention EOR. In this example, 80% of relevant content was missed using a normal keyword search.

Table 2 Results of queries applied to corpus

No. of Queries	Total results sampled	Δ Additional results found through QE	Standard Deviation
50	12,557	43%	0.338

In three specific queries, QE found no results. In four queries, keyword search found no results, but QE techniques did. An F-Measure was not calculated because the exact number of relevant results per query in the entire corpus was not known.

6.3 Number of Words in a Search Query

An interesting trend was observed when plotting the percentage of relevant documents for keyword search and KR QE. As the number of words (terms) in the query increases, relevant recall of keyword search begins to fall off compared to QE (Fig 3). The drop in the relative percentage of keyword to QE results is significant. When four query terms are used, QE techniques account for 70% of all *relevant* results found.

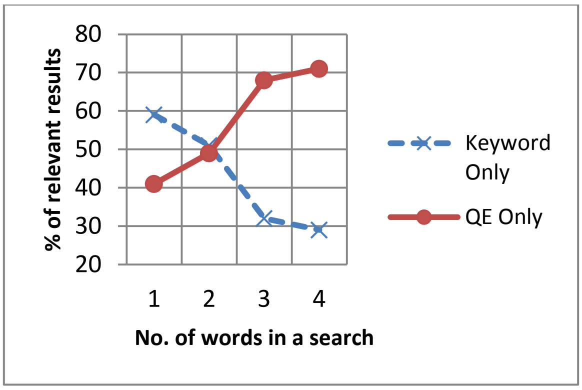

Fig. 3 –Proportion of all relevant results by method and by number of words used in a search

In Boolean keyword search this decline can be attributed to the Compounding of Semantic Field Ambiguity (CoSFA). Each word in a query has its own high precision semantic field, consisting of the range of linguistic possibilities described in Table 1. As the number of words in a search query increases, the number of possible combinatory semantic fields increases *exponentially* as does the possible linguistic variants. The latter is given by the product of the high precision Semantic Field Size (SFS) of each of the search words. This model explains why QE shows an increase in relative relevant results (to keyword) when moving from one to four search words as linguistic possibilities explode. It is highly improbable all member combinations will exist linguistically in a meaningful way within a query. When multi-word named entities and concepts are recognized in the search query, it will sometimes negate some of their constituent words' semantic fields. A model for total terms and linguistic variants used in QE for a query (q) is given by:

$$(P)SQ_n = (2^n) - 1$$

$$x = \{|(P)SQ_n \, \alpha \, (NE_q) \cup (MWC_q)|\}$$

$$y = \sum_{i=1}^{x} SFS_i \, \alpha(QKR)$$

$$z = SFS_1 \times SFS_2 \times \cdots SFS_x$$

Where:

n=Number of words in a search query

x=Total number of semantic fields realized in a search query

y=Total number of realistic query expansion terms

z=Total linguistic variants

(P)SQ= Powerset possible semantic fields in the search query

NE=Named entities found in the search query

MWC=Multi-word concepts found in the search query

SFS=Semantic field size

QKR=Quality of knowledge representation used in QE

The size of each semantic field is also variable. A concept like *Geophysics* (Discipline) will have a much larger SFS than a concept such as *Geophone* (Topic). The quality of the KR is critical, with the number of terms, coverage and modeling of variants in Table 1 all influence how many QE terms are used.

Realization of y into actual additional search results will depend on the nature of the corpus size and distribution. Putting these factors aside, in general, queries with more words will have more linguistic variants than queries with a smaller number of words. So as the number of query terms increases, CoSFA effects will also increase.

Using a geological example (although it is generic to any discipline) a user searching on *Limestone* will miss documents only mentioning *Chalk* (a type of *Limestone*). If a second word is added (a space meaning an AND) and a user searches on *Limestone Diagenesis* the user will miss documents only mentioning *Limestone Dissolution* (as *Dissolution* in this context, is a type of *Diagenesis*). However, they will also miss documents about *Chalk Diagenesis* and *Chalk Dissolution*. In other words, the combinatorial explosion of linguistic variants increases with the number of search words used.

In this study the large number of compound term concepts and artificially inferred synonyms and examples in the KR (Section 5.3) are playing their part in mitigating the effects of CoSFA. More research is required to model this effect.

6.4 Loss of Precision in Query Expansion

The loss of precision in QE (overall 81% to keyword 90%) can be explained by a number of factors. As expected, lack of Word Sense Disambiguation (WSD) is responsible for erroneous results. Homonymous acronyms will obviously need disambiguation. Even where non-homonymous acronyms were used, many were incorrectly matched, so non-unique (polysemy), even though they were believed to be unique (in the industry). This can be resolved by looking at the surrounding words, before matching in QE. This may increase precision by 7-14% (Schutze & Pedersen [25]).

The negative interference effects between semantic fields were not anticipated. A document containing the terms *Flooding* in the title and *H2S* as a keyword were parsed for QE. After parsing against the KR, the parent term *Chemical Compound* was added from the KR as a semantic relation to the document (H2S identified as Hydrogen Sulphide). When a test query was made on *Chemical Flooding* (an EOR technique) this document was returned, but it is not relevant. It was actually about Waterflooding (a different technique). H2S was a keyword relating to the byproduct of Sulphate Reducing Bacteria (SRB) in the process. Due to the ranking (Section 5.4), this result was low down in the ranking, but it was returned and it was not relevant. Using different proximity parameters for the search may mitigate these effects.

7. CONCLUSIONS

This paper has shown staff can miss, at a minimum, an average of 43% of valid results in a single Boolean keyword search. It has also shown that basic non-probabilistic techniques can improve Enterprise search effectiveness. When the text available to search is limited (often the case in an organization), the importance may be significantly increased.

It was discovered in keyword search the proportion of all relevant results returned decreases with the number of words used in a search as a proportion of all relevant results. This could be considered the *narrow search paradox*. Users add more terms to locate relevant information, but by doing so are proportionally missing more relevant information for their given search. This could be of wider significance, as evidence indicates users are adding more terms now to refine their queries than they did ten years ago, to cope with increasing information volumes. This may be less important for internet searching, but in a corporation critical information may remain undiscovered. To the author's knowledge this is the first time this phenomenon has been modeled in this context.

QE offers opportunities to mitigate the problems of missed information with keyword only search. This will improve business performance in a corporation without the need for any new technology. A strategy for query expansion should be part of any Enterprise search deployment, regardless of industry. Arguably, there are few reasons not to deploy non-probabilistic high quality knowledge representations as part of Enterprise Search. Theoretically, probabilistic methods offer even better results, providing information precision can be retained.

8. FUTURE WORK

Semantic networks, Natural Language Processing (NLP) and probabilistic techniques are being used to improve search recall. There are no published papers using a *relevant* oil and gas KR as a seed to these processes, particularly investigating increased recall whilst maintaining precision. During automatic extraction of concepts from public domain sources (Section 5.1), significant volumes of data were collected for concepts and associations. For each concept vector (C), with an occurrence frequency (CF), associated concepts (AC) were ranked by the frequency that association was found (FAC). Simply put, the relative probability (P) of a term being associated with the concept is FAC divided by CF.

For body text, a proximity weighting could be applied denoting the position within the document, i.e. (title, abstract), sentence, paragraph. Sub-headings on documents and font size in PowerPoint presentations could be used to weight. A function would need to be developed to mitigate artifacts caused by small concept occurrences. The Inverse Document Frequency (IDF) correction can be applied to decrease the importance of very common terms in the corpus. Given a sufficiently large and representative corpus, the resultant concept vectors *semantically define the concept itself*. The blending of a definitional KR with automatically identified concepts and relations anchored into a larger probabilistic semantic network needs to be tested in Enterprise search. An area of interest is the point at which semantic relatedness of a document or concept becomes too distant to show in an initial search result list or faceted breakdown, so these results or related concepts are presented differently as further options.

9. ACKNOWLEDGEMENTS

Thanks to everyone in Flare, Carel Skogstad and the multinational corporation whose documents were used in this study which cannot be named for confidentiality reasons. The Protégé tool from Stanford University was used for modeling.

10. REFERENCES

[1] IDC (2001), SmartLogic MindMetre (2011) Enterprise search user satisfaction surveys

[2] Hawking, D. (2004) Challenges in Enterprise Search, *CSIRO ICT, Conferences in Research and Practices in Information Technology*

[3] Chowdhury, S., Gibb, F., Landoni, M. (2011) Uncertainty in Information Seeking and Retrieval: A Study in an Academic Environment. *Information Processing & Management Volume 47 (2)(pp.157-175)*

[4] Gruber, T.R. (1992) A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition 5(2) (pp. 199-220)*

[5] The Energy Industry Profile of ISO 19115-1 (EIP)

[6] Ogilvie, P., Callan, J. (2001) The Effectiveness of Query Expansion for Distributed Information Retrieval. *In Proceedings of the 10th International Conference on Information and Knowledge Management (pp. 183-190)*

[7] Peng, J., Macdonald, C., He, B., Ounis, I. (2009) A Study of Selective Collection Enrichment for Enterprise Search. *In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2009*)

[8] Yi, X., Allan, J. (2009) A Comparative Study for Utilizing Topic Models for Information Retrieval. *31st European Conference on IR, LNCS 5478,(pp. 29-41)*

[9] Vorhees, E. (1994). Query expansion using Lexical-Semantic Relations. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 1994*.

[10] Kristensen, J. (1993) Expanding End User's Query Statements for Free Text Searching with a Search Aid Thesaurus. *Information Processing and Management Vol 29 No. 6 (pp. 733-744)*

[11] Segura, A., Salvador-Sanchez, Garcia-Barriocana, E., Prieto, M. (2010) An Empirical Analysis of Ontology-based Query Expansion for Learning Resource Searches using MERLOT and Gene Ontology. *Knowledge Based Systems Volume 24 Issue 1 (pp. 119-133)*

[12] Bhogal, J., Macfarlane, A., Smith, P. (2006) A Review of Ontology Based Query Expansion. *Information Processing and Management 43 (2007) (pp. 866-886)*

[13] Solskinnsbakk, G., Gulla, J. (2010) Ontological Profiles in Enterprise Search. *Journal of Data and Knowledge Engineering Vol 69 Issue 3 (pp. 251-260)*

[14] Resnik, P. (1999) Semantic similarity in a Taxonomy: An information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research 11 (pp. 95-130)*

[15] Quillian, M. (1968). Semantic Memory. *Information Processing (pp. 227-270) MIT Press.*

[16] Collins, A., Loftus, E. (1975) A Spreading-activation Theory of Semantic Processing. *Psychological Review Vol. 82 No. 6., (pp. 407-42*8)

[17] Crestani, F. (1997) Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review Volume 11 Issue 6.*

[18] Wojtinnek, P., Pulman, S. (2011). Semantic Relatedness from Automatically Generated Semantic Networks. *Proceedings of the 9th International Conference on Computational Semantics (IWCS11) (pp. 390-394)*

[19] Liu, W., Weichselbraun A. Scharl, A. (2005). Semi-automatic Ontology Extension Using Spreading

Activation. *Journal of Universal Knowledge Management and Proceedings of iKNOW 2005*.

[20] Velardi, P., Navigli, R., Faralli, S. Martinez, J. (2012). A New Method for Evaluating Automatically Learned Terminological Taxonomies. *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012), May 21-27, 2012*

[21] Woon, W. and Madnick, S. (2009) Asymmetric Information Distances for Automated Taxonomy Construction. *Knowledge and Information Systems Volume 21, Number 1 (pp. 99-111)*

[22] Vorhees, E.M., Harman, D. (1998) TREC-7 Experiment and Evaluation in Information Retrieval

[23] Taghavi, M., Patel, A., Schmidt, N., Wills, C., Tew, Y. (2011). An Analysis of Web Proxy Logs with Query Distribution Pattern Approach for Search Engines. *Computer Standards and Interfaces (pp. 162-170)*

[24] Shiri, A., Chambers, T. (2009) Information Retrieval from Digital Libraries: Assessing Potential Utility of Thesauri in Supporting Users Search Behavior in an Interdisciplinary Domain *Proceedings of the 10th International Conference of the International Society for Knowledge Organization (ISKO)*

[25] Schutze, H., Pedersen, J. (1995) Information Retrieval based on work senses. *In Symposium on Document Analysis and Information Retrieval. In Proceedings of SDAIR'95 Las Vegas Nevada (pp. 161-175)*