Improving Enterprise Wide Search in Large Engineering Multinationals: A Linguistic Comparison of the Structures of Internet-Search and Enterprise-Search Queries

David Jones, Yifan Xie, Chris Mcmahon, Marting Dotter, Nicolas Chanchevrier, Ben Hicks

Improving Enterprise Wide Search in Large Engineering Multinationals: A Linguistic Comparison of the Structures of Internet-Search and Enterprise-Search Queries

David Edward Jones[1], Yifan Xie[2], Chris McMahon[1], Marting Dotter[2], Nicolas Chanchevrier[2], Ben Hicks[1]

[1] Department of Mechanical Engineering, University of Bristol, UK
dj13730@bristol.ac.uk, Chris.McMahon@bristol.ac.uk, bh13105@bristol.ac.uk

[2]Airbus Group, Toulouse France
yifan.y.xie@airbus.com, martin.dotter@airbus.com, nicolas.chanchevrier@airbus.com

Abstract. Understanding how users formulate search queries can allow the development of search engines that are tailored to the way users search and thus improve the knowledge discovery process, a key challenge for Product Lifecycle Management (PLM) systems.

This paper presents part-of-speech (POS) statistical analysis on two sets of 'Top 500' search query lists in order to compare Internet search with enterprise search with the aim of understanding how enterprise search queries differ from Internet search queries. The Internet queries were obtained from the keyword research company WordTracker.com and covers the month of January 2015. Enterprise search logs were obtained from a large multinational engineering organization and represent the first six months of 2014.

The results show enterprise search users are far more likely to search using nouns, with 97% of queries containing at least one noun. This compares to 89% for Internet users. 60% of enterprise queries are single nouns compared to 38% for Internet search users. In total, enterprise queries fell into 41 lexical classes (noun-noun/adjective-noun/etc.) whilst Internet search contained 95 classes. Of those 41 classes only 12% contained no nouns, compared to 21% for Internet search. 80% of the enterprise search queries can be covered by just four Lexical classes compared to 15 for Internet search. 90% coverage required 11 classes for enterprise and 44 classes for the Internet.

These findings appear to support existing literature in that they show a preference for enterprise searches for specific information using domain specific terms. This paper concludes by considering the implications of these findings for enterprise search systems and PLM in the context of a large engineering organization and in particular proposes two areas of future research.

Keywords: Knowledge Management, Enterprise Search

1 Introduction

Enterprise wide search systems are central facets of knowledge management and the primary means for finding and re-finding information across the product lifecycle. This is particularly true for large multinational engineering organizations where people, information and expertise are dispersed across multiple sites and multiple countries. Many of the tools and techniques employed in enterprise or intranet search were originally developed for Internet search and while users expect the same level of results as Internet search, their opinion of intranet search performance is that it often falls short of Internet search [1, 2].

In this regard, there are comments in the literature on the difference between Internet and intranet search systems, and specifically how users of intranet search expect the quality of results offered by Internet search and are commonly disappointed with the state of the art enterprise systems on offer. For example, in a small scale qualitative study reported by [2] into the usefulness of enterprise search using Microsoft SharePoint 2013 in an automotive engineering company, research found issues with users being able to formulate queries for the required results; users having difficulty in extracting information from the range of document types; inconsistent usage of metadata and also the "...misleading built-in relevance model of the enterprise search engine." that leads to poor ranking of search results.

While both Internet and intranet search systems deal with finding information, the differences between the two are important and must be studied and understood if the utility of enterprise search is to be comparable to that of Internet search. It could quite possibly be the case that some of the solutions to improved intranet search lie in the aspects that make them different rather than those that are common.

To date, work in the area of improving enterprise search has focused on three main areas: building knowledge organizational schemes (taxonomies and ontologies), personalized search using user characteristics and faceted search. Each of these aims to improve search by applying structure to the dataset to make it more straight-forward to process and use. Taxonomies capture the connection between terms and represent domain data in a tree structure and ontologies capture the relationship between terms and represent these in a network like structure [3]. Personalized search attempts to understand the user and, through this, the context of a search, for example, a member of a finance team is more likely to be interested in finance related documents while a member of a design team is more likely to be interested in engineering related documents [4]. Faceted search stores the dataset in a number of faceted classifications, effectively multiple taxonomies, that allows the navigation of the dataset through these facets which can help to meet the different perspectives of users [5, 6].

One area that has seen some limited investigation in Internet search but to date has not been seen in the field of intranet search, is that of linguistic analysis of search query logs [7-9] and in particular, a comparison between how Internet and intranet users construct their search queries. Understanding how queries are structured can be used in both the term extraction process during indexing to improve precision of results returned by the search engine [10] and in devising strategies for facetted classification and/or taxonomies.

Linguistic analysis of search logs involves the parsing of queries through a part-of-speech (POS) tagger. POS taggers parse text and tag each word with its lexical category or parts of speech class (e.g. Noun, Verb, Adjective, etc.). The goal of such analysis is to align how users phrase queries with the term extraction process and optimize the precision of results returned. Nakagawa in [11] states that 85% of domain specific terms are said to be compound nouns and uses this to improve the extraction of domain specific terms using a combination of POS tagging to identify compound nouns and statistics.

In a similar manner to Nakagawa in [11], this paper presents a comparison of Internet and intranet search queries to better understand what makes intranet search different to Internet search within a large engineering organization. Following a detailed discussion of the results it then considers the implications of the findings for improving enterprise search over the product lifecycle and within the context of PLM systems**.**

2 Method

This section is divided into two subsections. The first discusses the data obtained for the investigation and the second discusses the technique and tools used for part-of-speech tagging.

2.1 Data

Obtaining accurate Internet search engine query logs is a relatively difficult task with the large search engine giants only providing limited access to top-n (n<25) results at most. Hence, a 'Top 500' search query list was obtained from WordTracker.com, a company specializing in keyword data collection that provides third parties with an API, Keyword Research tool and Reports for the exploration of this data for purposes such as search engine optimization. WordTracker.com provided a global Top 500 query report for the month of January 2015. The top 10 results from this set are shown in Table 1. Intranet search query logs were provided by the Airbus Group and comprise the top 500 queries submitted to their Business Search tool. Data was collected from January 1st 2014 through to June 30th 2014 and covers nearly 1.1 million searches with approximately a third of those being unique and executed by more than 68,000 unique users.

Table 1. Top 10 Internet and Intranet Search Queries and Search Frequency

Internet Search Queries		Intranet Search Queries	
Query	Frequency	Query	Frequency
youtube	9924821	docmaster	8736
movies	8721604	icc	7186
facebook	8085544	lexinet	7022
google	6968440	webex	7012
entertainment	6067158	pwinit	6591

Internet Search Queries	
Query	Frequency
search	5186360
craigslist	4888389
kinox	4828994
hood stars clothing	3735957
download	3006655

Intranet Search Queries	
Query	Frequency
uvisit	3982
airnav	3310
eds	2967
zamiz	2766
edms	2692

2.2 Part of Speech Tagging

Python's Natural Language Toolkit (NLTK) provides an off-the-shelf POS tagger that automatically parses text and tags words with their lexical categories or parts of speech (noun, adjective, verb, etc.). For the purposes of this work, the default NLTK POS tagger in NLTK version 2.0b9 and Python version 2.7.6 were used. Terms from both datasets were parsed by the tagger one at a time and the resultant tagged term set returned. Table 2 shows a list of all possible individual POS tags. Where queries contain more than one word both words are tagged, for example, *'aeroplane wing'* would return (*'aeroplane* NN), (*'wing'*, NN) - a noun-noun (NN NN) bigram. For the purposes of this paper, a combination of POS tags will be referred to as a Lexical Class.

Table 2. List of POS Tags and their Corresponding Description

POS Tag	Description
CC	coordinating conjunction
CD	cardinal number
DT	determiner
EX	existential there
FW	foreign word
IN	preposition/subordinating conjunction
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
LS	list marker
MD	modal
NN	noun, singular or mass
NNS	noun plural
NNP	proper noun, singular
NNPS	proper noun, plural
PDT	predeterminer
POS	possessive ending
PRP	personal pronoun
PRP$	possessive pronoun
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	particle
TO	to
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	verb, gerund/present participle
VBN	verb, past participle
VBP	verb, sing. present, non-3d

POS Tag	Description
VBZ	verb, 3rd person sing. present
WDT	wh-determiner
WP	wh-pronoun
WP$	possessive wh-pronoun
WRB	wh-abverb

4 Results

Figure 1 and Figure 2 show the most frequent Lexical Class frequencies of the Airbus Business Search and WordTracker.com Internet search top 500 queries respectively. Comparing the two graphs, the most obvious differences between the two sets of data is the variety in different lexical classes: Business Search contained 41 different classes while the WordTracker.com dataset contained more than double the classes at 94. Figure 3 combines the most frequent queries from both data sets and shows the most popular lexical class for Internet and intranet are single nouns. Business Search contains over a third more single noun queries than Internet search with 60% business queries being single nouns compared to 38% of Internet queries. The Internet queries contain twice as many plural nouns with 10% compared to 4% for intranet. For noun-noun bigrams, the figures are closer with 10% for business and 8% for Internet. The final significant result to mention is the percentage coverage per number of lexical classes, 80% of the business search queries are covered with just 4 lexical classes and 90% coverage is achieved with 11 classes, these are far fewer than the Internet queries where 15 classes are required to reach 80% and 44 classes to reach 90%. An important note is those 4 lexical classes are all nouns: singular nouns, noun-noun bigrams, proper nouns and plural nouns. Expanding this to the full set of queries, 97% of business search queries contain nouns compared to 89% for Internet queries.

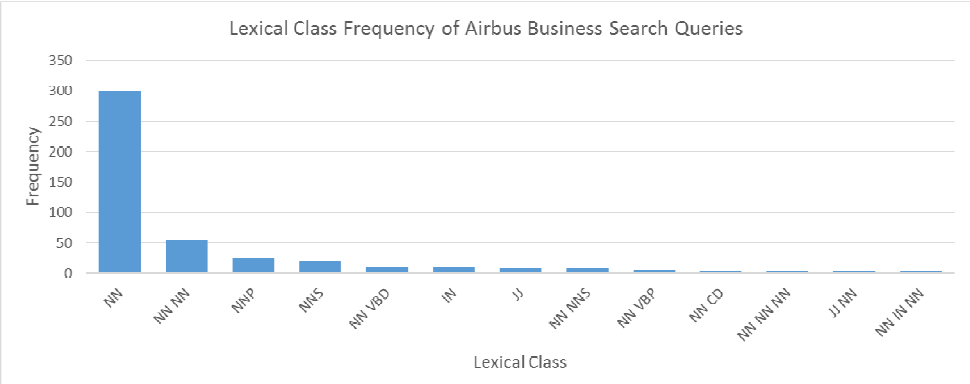

Figure 1. Lexical Class Frequency of Airbus Business Search Queries

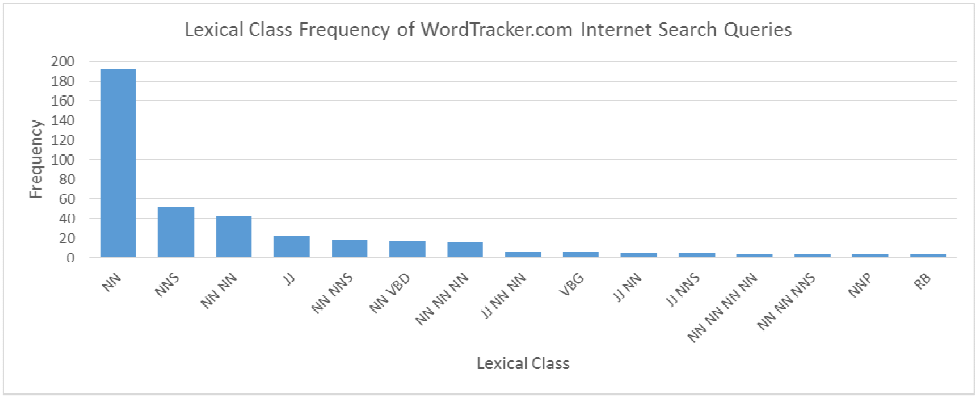

Figure 2. Lexical Class Frequency of Internet Search Queries

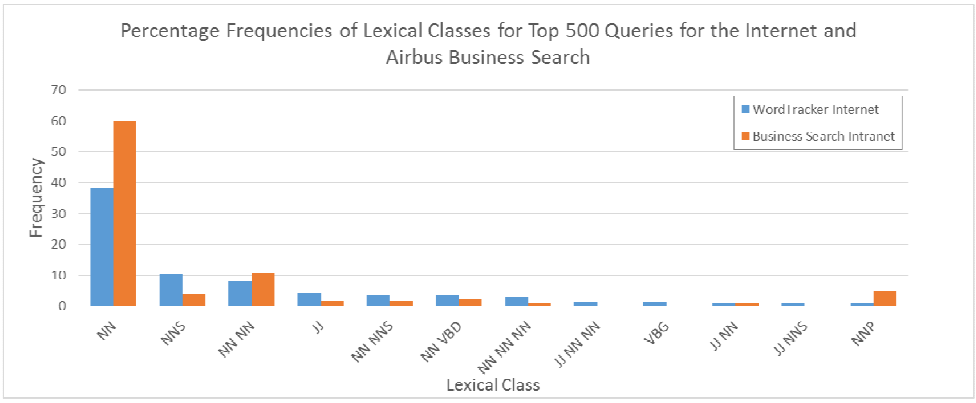

Figure 3. Percentage Frequencies of Lexical Classes for Top 500 Queries for the Internet and Airbus Business Search

4 Discussion

In the comparison of Internet and intranet search queries from a large engineering organization it has been shown that there are some distinct differences between the two. In summary, the differences show that intranet search queries are far more noun based with less lexical variety in the way users construct their queries. The remainder of this paper will discuss the implications of these findings within the context of PLM and enterprise wide search.

[1] states the during intranet search users are more specific about their search requirement and frequently search for documents they know exist, in the case of intranet search a good result is generally perceived as the result with the right answer. The findings presented here could be interpreted to confirm this; the higher use of nouns within intranet search can be explained by the fact that Airbus contains a high number of explicit Applications, Documents, Process, etc. and that users are searching for these rather than using more general textual descriptions. As an example, the first

two non-noun queries in the top 500 Internet search queries are '*2015*' (classified as a cardinal number rather than the name of a year) and 'generic' compared to '*unified planning'* for the intranet queries.

To explore this result further and the proposition that nouns are more likely in intranet search and that they relate to business systems and operations, the business search queries have been classified by an Airbus user group. Table 3 shows the results from the classification of the top 574 Business Search queries by Airbus staff and influenced by the set classes outlined in [12]; each query can belong to multiple classes. Incidentally, [12] discusses the development of a context based search platform at EADS ((European Aeronautic Defence and Space) formally the parent company of Airbus and has now been rebranded as the Airbus Group) and the classes highlighted are used to represent search context. Of the top 574, 85 were classed as Unknown and the highest top 5 classes were *Applications*, *Documents*, *Activities/Processes*, *Organization* and *Product* and these classes cover 78% of business search queries. This list again confirms that intranet search users are predominantly searching for specific business related information.

Table 3. Intranet Search Queries Classified by Airbus Users

Class	Frequency
Application	172
Document	108
Activity/Process	99
Organization	81
Product	80
Project	25
Role	23
Devices	19
Discipline	17
Gate	2
Member	2
Unknown	80

The question now is how does all this apply to PLM and improving enterprise search? The results have shown that users search for real-world, business related 'things', things that are specific to the Airbus domain. The process of generating search indexes, whether Internet or intranet, is to extract all 'meaningful' terms from a document and index each document against the terms it contains. This works for the Internet as everything is required to be searchable by anyone within any context but for intranet search, if we can say that users are searching for domain specific things, then we can hypothesise that the index does not need to contain terms outside of a list of domain specific terms – a domain specific index. Removing unnecessary terms from the index can cut down the noise in the data set and improve precision.

In addition to smaller indexes, once a list of domain specific terms is obtained the indexing process can begin to move beyond pure term extraction. The challenge becomes more akin to those addressed by the field of machine learning where techniques like classification, multifaceted classification and case-based reasoning automate the process of identifying relationship and similarities between documents based on the characteristics of the document. This would for example result in a more intelligent understanding of what makes a document about *WebEx* a document about *WebEx* using additional meta-data (author, date of creation, location (stored) and

(created) for example). This would lead to the creation of more intelligent search systems returning results of higher relevance.

The results also confirm that strategies to improve intranet search such as generating taxonomies and ontologies which add structure to data and attempt to 'understand' the context and relationships of information within a domain are entirely appropriate. This would help to align how search indexes are generated with how users approach their searches.

The future of enterprise wide search requires domain specific search indexes that are specific to the user requirements, well-structured and provide a higher precision of results over the range of results returned. A system based on these attributes also opens the door to reinventing the front end of search engines. [13] Introduces a strategy for artefact-based information navigation, a system where documents are navigated within a visual representation that captures the context of the search. A web-based 3D Formula Student racing car and student reports are presented but the approach is extendable to data relating to other physical artefacts. The user manipulates the model to locate the area of the object of interest. Documents are represented in the model as Points-Of-Interest (POI). Looking at a POI generates a Google style list of results. There is no reason why the top five query classifications from Table 3 (*Applications*, *Documents*, *Activities/Processes*, *Organization* and *Product*) could not be visualized in such a way and indexed in the method proposed above. Figure 4 is an example of what such a system could look like, with an Airbus A380 representing the Product class.

Figure 4. Example a Product Artefact-Based Information Navigation System

Taxonomies and Ontologies are in essence textual representations of real world relationships between objects and so the visualization of the classifications in Table 3

in the manner depicted in Figure 4 has the added advantage of showing these relationships in a way that is more akin to the real world. For example, it is possible to see that the *wing* connects to the *fuselage* and comprises of *fairings*, *flaps*, *ailerons* and *nacelles* which in turn connect to the *engines* and so on. The representation of information in this way could improve the way engineers find information and discover new knowledge as they align the search system with the visual and functional nature that is inherent in the engineering process, product architecture and the design representations used.

In terms of the method employed, the accuracy of the POS tagger will impact on the results. Similar work outlined in [8, 9] take time to focus on improving the accuracy of the POS tagger within the domain that they operate. The work presented here deliberately used an off-the-shelf POS tagger and treated each list of queries equally rather than attempt to improve the accuracy for both and then attempt a comparison. The first non-noun query '*unified planning*' is grammatically a non-noun query but in reality is an Airbus system and therefore could arguably be treated as a single noun (a similar example from Internet search would be '*hood stars clothing*' – an organisation).

5 Conclusion

The paper compared the way user's structure Internet and intranet search queries in an attempt to better understand the difference between the two types of search and ultimately improve intranet search. Literature has shown the usability and quality of intranet search to be lacking when compared to Internet, and that intranet search users require a higher level of precision from a search system rather than the balance of precision and recall provided by Internet search. The results presented here go some way to verify these findings and reveal that:

1. Intranet users within Airbus are more likely to phrase their queries using noun, with 97% of search queries containing nouns (compared to 89% for Internet queries) and use far less variety in how queries are formulated, with intranet queries falling into 41 lexical classes with just four of those required to cover 80% of queries, compared to 94 for Internet and 51 to cover 80% of queries.
2. The intranet queries could be classified into distinct business related classifications. The top five of which are Applications, Documents, Activity/Process, Organization and Product and these top five represent 78% of business search queries.

This paper concluded with a discussion on the implications of these findings in the world of PLM and summarized that the current strategies of adding structure around search index terms appears to mirror the way users structure queries. Based on this and the observation that users search with domain specific terms, two areas of future research are highlighted.

1. The investigation of the creation of domain specific search indexes with machine learning techniques like classification and case-based reasoning

being used to generate more intelligent search indexes than those created by pure term extraction alone.

2. Changing search interfaces to represent the information search space via a visual representation such as product, process or organizational structure. Further a number of visual interfaces could be combined to support visual-multi-faceted search and/or support different users/perspectives.

Acknowledgments. This research is funded via an EPSRC CASE AWARD, the Language of Collaborative Manufacturing (LOCM) Project (EPSRC grant reference EP/K014196/1) and the Airbus Group. The Authors would like to thank colleagues at Airbus and the University of Bristol for support and contribution.

References

1. Mukherjee, R. and J. Mao, *Enterprise search: Tough stuff.* Queue, 2004. **2**(2): p. 36.
2. Stocker, A., et al. *Is enterprise search useful at all?: lessons learned from studying user behavior*. in *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*. 2014. ACM.
3. Varma, V., *Use of ontologies for organizational knowledge management and knowledge management systems*, in *Ontologies*. 2007, Springer. p. 21-47.
4. Hawking, D., et al. *Context in enterprise search and delivery*. in *Proc. IRiX Workshop, ACM SIGIR*. 2005.
5. McMahon, C., et al., *Waypoint: an integrated search and retrieval system for engineering documents.* Journal of Computing and Information Science in Engineering, 2004. **4**(4): p. 329-338.
6. Sacco, G.M., *Dynamic taxonomies: A model for large information bases.* Knowledge and Data Engineering, IEEE Transactions on, 2000. **12**(3): p. 468-479.
7. Allan, J. and H. Raghavan. *Using part-of-speech patterns to reduce query ambiguity*. in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 2002. ACM.
8. Barr, C., R. Jones, and M. Regelson. *The linguistic structure of English web-search queries*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2008. Association for Computational Linguistics.
9. Ganchev, K., et al. *Using search-logs to improve query tagging*. in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. 2012. Association for Computational Linguistics.
10. Hulth, A. *Improved automatic keyword extraction given more linguistic knowledge*. in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003. Association for Computational Linguistics.
11. Nakagawa, H. and T. Mori. *A simple but powerful automatic term extraction method.* in *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14*. 2002. Association for Computational Linguistics.
12. Redon, R., et al., *VIVACE context based search platform*, in *Modeling and Using Context*. 2007, Springer. p. 397-410.
13. Jones, D.E.C., Nicolas; McMahon, Chris Hicks, Ben. *A Strategy for Artefact-Based Information Navigation in Large Engineering organisations (InPress)*. in *ICED15: The 20th International Conference on Engineering Design.* . 2015. Milan, Italy.