A Case Study of Utilising Concept Knowledge in a Topic Specific Document Collection

Gavin Shaw & Richi Nayak
Data Sciences Discipline, Science & Engineering Faculty
Queensland University of Technology
2 George St, Brisbane, 4000, Queensland, Australia
gavin.shaw@qut.edu.au; r.nayak@qut.edu.au

Abstract

The use of 'topic' concepts has shown improved search performance, given a query, by bringing together relevant documents which use different terms to describe a higher level concept. In this paper, we propose a method for discovering and utilizing concepts in indexing and search for a domain specific document collection being utilized in industry. This approach differs from others in that we only collect focused concepts to build the concept space and that instead of turning a user's query into a concept based query, we experiment with different techniques of combining the original query with a concept query. We apply the proposed approach to a real-world document collection and the results show that in this scenario the use of concept knowledge at index and search can improve the relevancy of results.

Keywords: Text Mining, Document Concepts, Term to Concept, Concept Search, Case Study, Wikipedia.

1 Introduction

Text mining is a critical function used to discover documents that are related to a user's query of interest. This is usually facilitated by searching an index via a query and matching the terms from a query with those contained within documents. Such searching and mining activities can be enhanced through a variety of techniques at both ends of the process; document indexing time and/or query time. Given that documents can contain a large volume of text and many differing terms or be short with much fewer unique terms; all within the same collection; traditional term searching and matching can yield poor results [Fang, 2004; Lv, 2011]. The keyword-based text matching methods can miss documents that are relevant but use different terms.

One popular approach to improving performance of simple text matching approaches is to attempt to discover, add and utilize concept level knowledge within the document set and user queries. The identified concepts operate at a higher information space and reduce the number of 'terms' associated with a given document. A key advantage to the use of concepts, aside from providing a set vocabulary is that they can result in documents being brought together under one concept even when they use different terms/words to describe the same concept. Furthermore, if a 'concept only' search can be conducted, the costs of handling user queries can be reduced, although there may be a trade-off in the quality of the results due to the more 'generalised' nature of concepts.

This paper examines and evaluates the improving of document search retrieval for an industry dataset that has the potential to play an important role within a business unit of a much larger company. The document collection was gathered to help the unit undertake a foresight and future options development and help analysts looking for information relevant to strategic risks and/or structural changes. Thus ensuring highly relevant results are returned and ranked highly is key to supporting these activities.

There are several key points and contributions in our work;

1. Utilizing a publicly available, general purpose (eg. not domain specific) information source, Wikipedia, to build a domain topic specific concept space instead of purpose building a custom concept space; which can be resource demanding.
2. In using said general purpose knowledge source, instead of using all or a random selection of its content; to generate term to concept links, we use a specific portion that is identified as relevant to the particular application domain and control how far we crawl to build the concept space and associated term to concept links.
3. Instead of taking a user's query and converting it to a concept only query; via the built concept space model; such as that proposed in [Egozi, 2011]; we create a hybrid query by combining the traditional term/phrase based search with a concept based search. Similar to the traditional approaches the concepts can be discovered by mapping the user's query terms to concepts via the concept model, but can also be drawn from the documents returned as results to an initial query formed from the user query. Further, these two approaches to building the concept component of the hybrid query can be used together or separately.
4. For this work we test these ideas on a real world industry document collection gathered by a major business operating in the financial sector of Australia and evaluate the performance of such a concept space and hybrid query approach.

To our knowledge the specific combination of how the concept space was built and utilized is new, further its evaluation against industry data shows its potential for

use outside research, given that previous proposed approaches; such as those in [Milne, 2008; Medelyan, 2008]; often do not get evaluated on a dataset from the commercial world, should make the approach of interest to industry.

The rest of the paper is as follows; Section 2 reviews related background materials, while Section 3 outlines the proposed method for building a concept space and the utilizing it at index and query time. Section 4 presents our experiments, results and evaluations of the proposed approach. Section 5 concludes our work and the paper.

2 Background & Related Work

The key challenges of using concepts in text mining is identifying a list or vocabulary of concepts, determining which concepts a document is related/relevant to and then taking advantage of this extra information when a query is submitted. The first challenge, identifying concepts, has received much attention [Hou, 2013; Egozi, 2011, Huang, 2009; Medelyan, 2008; Mihalcea, 2007]. A common and popular source for determining concepts is Wikipedia (although WordNet and BabelNet have also been considered), where each article represents a concept [Hou, 2013; Huang, 2009; Medelyan, 2008; Mihalcea, 2007] and text that refers to a given article becomes terms/phrases that represent that concept. With approaches that use Wikipedia, either the entire articles collection is used to build a concept list, or a random subset is used. This approach is fine for document collections that cover multiple topic areas, but when you have a document collection focused on a much smaller topic area, making concepts from unrelated areas available for possible use is likely to lead to issues with performance at query time.

However, a purpose built concept space, for a specific domain/topic can be impractical depending on the scope and the level of detail desired and can involve the need for domain experts to build such a space. Thus it is of interest to determine and discover whether such a focused concept space can be built from an existing, much broader space without being 'polluted' by unrelated domains. There does not appear to be much existing work focusing on this.

For the second challenge, works such as [Medelyan, 2008; Mihalcea, 2007; Milne, 2008] outline approaches to determining how to relate documents to concepts. They include identification of term to concept mapping [Medelyan, 2008]; selection of key text to link [Mihalcea, 2007; Milne, 2008] and identification of the most relevant concept to a piece of key text when there are multiple possibilities [Medelyan, 2008; Milne, 2008].

The final challenge, using concepts at query time, usually involves some form of query enhancement or expansion. Recent work in [Carpineto, 2012] looks at many automatic query expansion applications; however none of them involved dealing with concepts. Other works identified that present concept oriented works do not explicitly state how such knowledge would be used at query time to improve result performance. Work in [Egozi, 2011] outlines how a query is converted into an Explicit Semantic Analysis concept vector which is used to find the best matching documents. This approach however results in the search becoming concept only, which can work in some applications, but not all domains, applications and queries. The weak point with an approach, such as that in [Egozi, 2011]; or [Huang, 2009; Medelyan, 2008; Mihalcea, 2007; Milne, 2008] if concept only querying is used; is that when a user's initial query terms cannot be successfully mapped to one or more concepts for search then the query will fail. Conversely if the initial terms never take advantage of being mapped to concepts, then this extra, higher level knowledge is unused. Further still, each approach with their limitations, are unlikely to always be able to substitute for the other.

Given these limitations and the importance of this document collection to the industry partner, it is essential to ensure that the proposed approach be able to deal with the widest possible range of queries and successfully return relevant results. Thus in order to achieve this we will combine the traditional term/phrase searching with a concept based searching to have a hybrid query search approach to finding relevant documents.

3 Proposed Approach of Concept Space Generation and Querying

In this section we outline our proposed approach to enhance text mining through concept discovery & mapping, and query searching. There are three main components; discovery of relevant concepts and their mapping to terms (eg. building the concept space); document concept discovery and selection; and finally, query concept discovery & utilization for search.

We start this section by first introducing important definitions related to our proposed methods, followed by presenting each of the three main components in turn.

3.1 Definitions

The following are key definitions used in our proposed approach.

Definition 1 – Important Term: An important term is a single word or n-gram (composed of successive words) found within a document, from the collection that matches exactly with one or more hyperlink text entries, obtained from the term to concept mappings discovered via Wikipedia. The important term thus maps to one or more Wikipedia articles, which using their titles, represent concepts potentially relevant to the document.

Definition 2 – Concept: A concept is generated by the presence of a Wikipedia article page (not counting disambiguation, red links, category, 'etc' pages) discovered during the crawl of Wikipedia. The concept is represented by the title of said Wikipedia article page. Thus, a Wikipedia article can be considered as a concept, identified by its title.

Definition 3 – Context Concepts: A given document usually contains many terms/n-grams that potentially link to multiple concepts. However, there are a handful of terms/n-grams that only link to one concept each. The concepts which have no competing candidates for the term/n-gram that mapped to them can be known as context concepts for the given document. These particular

concepts help to describe the document and set the context in which it is present; hence the name. Context concepts can be used to help determine which concept, from a list of candidates, is the most relevant/related to the given document for a given term/n-gram.

3.2 Concept Space Generation

In order to be able to enhance the document index and user queries with concepts, it is first necessary to discover concepts that would be relevant to the topic(s) covered by the document collection and the important terms that are used to refer to them.

One difference between our use of Wikipedia and that of other works is that we use a targeted subset of the Wikipedia collection as opposed to the entire collection or a random sample that has been used in previous works. In our case the targeted starting point for our subset of Wikipedia is the category page for 'Finance'. We have taken this approach due to the nature of our document collection and industry partner. They are interested in the financial sector and thus tagging documents with concepts from other categories/topic areas, such as Science Fiction, Cooking, Anime etc, is unlikely to be relevant. Thus using the whole Wikipedia collection for concept discovery would lead to the inclusion of concepts which have none or very little relevance to the area of finance and the expected user queries.

Similarly, random selection of Wikipedia articles is also likely to lead to a similar situation in which irrelevant concepts are linked to the documents. Further, the use of a set of random articles may be worse as the selected set may not include any articles and hence concepts, in the topic area(s) related to finance.

The following figures (1, 2 & 3) demonstrate at a high level the approach to build the topic specific concept space, with associated term/phrase to concept mappings, from the full Wikipedia resource.

Figure 1 outlines the overall Wikipedia crawling process, where we start at a specific page and determine whether it is a category page (Figure 2) or an article page (Figure 3) and process it accordingly. After all desired pages have been processed we have a full listing of term to concept mappings for which we calculate a commonness score, using Equation 1. Finally, the term list, concept list and term-concept list are brought together to build the concept space that will be used during indexing and querying/searching.

The commonness score is determined via the following:

$$Commonness(t,c) = \frac{count(t \mid c)}{count(t)} \qquad \text{Eq. 1}$$

where t is the term/n-gram, c is the concept, $count(t|c)$ is the number of times text t was discovered to link to concept c and $count(t)$ is the number of times term/n-gram t was discovered to link to a concept, including c.

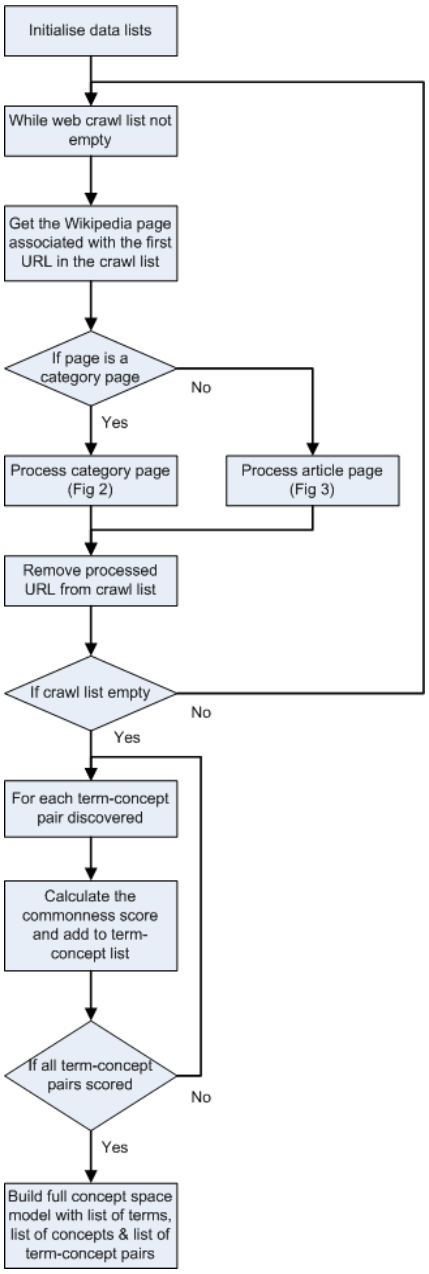

Figure 1. Overall concept space generation approach.

Figure 2 outlines the approach with processing category pages from Wikipedia. These pages do not form concepts, but instead provide a list of subcategories and articles (which are concepts) that fall under the category. For category pages the process is fairly straight forward, the current crawl depth of the page from the staring page in Wikipedia (eg. the number of hyperlink hopes) is compared against the maximum distance allowed. If it is less then all of the subcategories present are added to the crawl list with a crawl depth one greater than the crawl depth of the current category page. If the current crawl depth equals or is greater than the maximum allowed then the subcategories are not to be processed and the links to them are not added to the crawl list. Finally all of the links to article pages (eg. concepts) are added to the crawl list for processing.

We place a limit on the crawl depth to stop the process of generating the concept space from attempting to include a large portion of Wikipedia, which would then introduce unrelated concepts into the model. The

maximum crawl depth keeps the concepts collected more closely related to the original starting point and thus the specific topic of the domain remains relevant to this document collection.

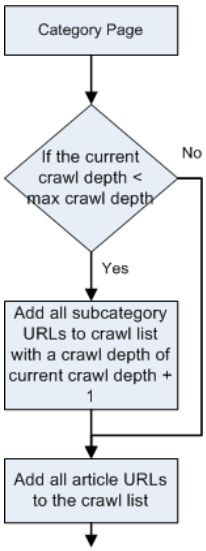

Figure 2. High level flow for processing category pages from Wikipedia.

Figure 3 outlines the approach for processing article pages from Wikipedia, which will form the basis for concepts in our concept space. For each article page all of the hyperlink text and their target URLs are extracted from the article's body. The last part of the target URL specifies the Wikipedia page and corresponds to the title of said page. This thus becomes the name of the concept that will be used within our concept space. With this we are able to build term/phrase to concept mappings. To support the concept space being built we maintain a list of terms, a list of concepts and a list of term-concept pairs, with every entry in these lists having an associated occurrence frequency. The text to URL titles extracted from each article page is added to these lists to build the concept space. Once all of the extracted hyperlink text and target URL pairs are processed, we have finished with the article page.

For the list of concepts it is important to note that it is not just the name of the concept (article linked to), but also a complete list of term/phrase to concept mappings of the hyperlinks from within this target article is included.

3.3 Document Concept Mapping

Once the concept generation via Wikipedia is complete, it is then necessary to index the document collection with the proposed enhancements. In our approach to this document collection, we implemented two main enhancements over the basic standard of indexing a document's title and textual contents. The enhanced index includes separate entries for a document's important terms and its associated concepts.

As the document concept mapping is not the primary focus of our proposed approach (rather the concept space generation and query enhancement are) and to save space, we do not go into great detail. Suffice to say, the approach involved takes each document, tokenises its content and then process each term in the following way. If the term is not a stopword and is found in the term to concept list previously developed then it becomes an important term and a set of candidate concepts for that term for the current document are identified. This is also done for n-grams up to the desired size. After a document's text has been processed and mappings to potential concepts identified, then the most relevant concept for each important term is identified. If an important term only maps to a single candidate concept then that concept is associated with the document and used as part of the context for the document.

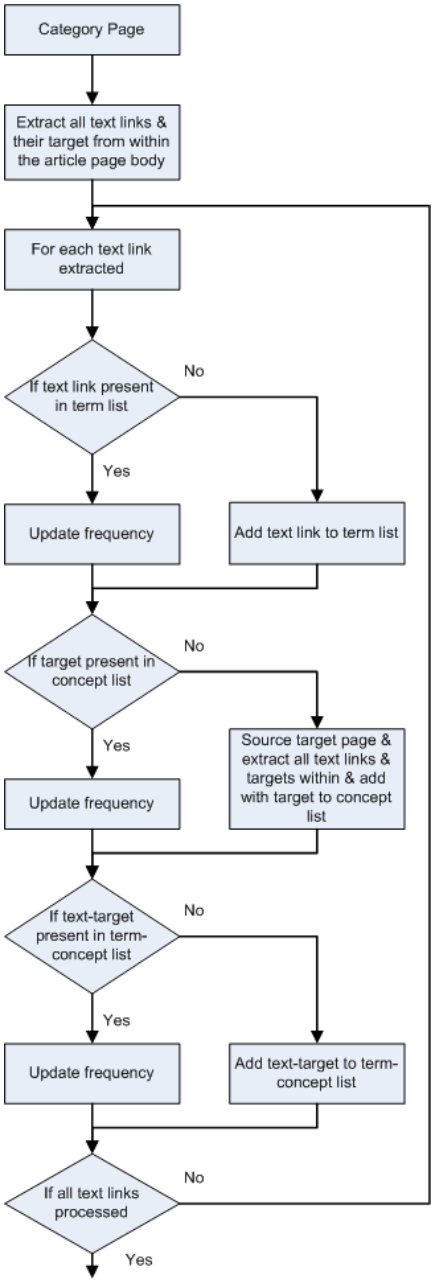

Figure 3. High level flow for processing article pages from Wikipedia.

For important terms that have more than one candidate concept we calculate the relevance scores of each candidate concept to the document for that term, using Equations 2 & 3 and the context concepts. The candidate concept with the highest relevance score for the important term is the one associated with the document. This

calculation and association happens for each important term instance.

Finally, once the most relevant concept for each important term is determined the top-n important terms and the top-n concepts; measured by frequency; are identified and added to an associated field for the document to be added to the index. The full listing of important terms and concepts are maintained in separate fields, giving the option of performing searches against the top terms or concepts or the full set of terms or concepts for each document. This supports greater searching options and allows us to test the performance of using both a document's full concept set and separately a document's set of top (most frequent) concepts.

As required by the approach taken to map documents to concepts prior to indexing, the following equations are used to determine the relevance of a candidate concept for a given term/n-gram for a given document.

$$SIM_{C1,C2} = 1 - \frac{(\max(\log(|C1|), \log(|C2|))) - (\log(|C1 \cap C2|))}{N - (\min(\log(|C1|), \log(|C2|)))} \quad \text{Eq. 2}$$

where *C1* and *C2* are concepts for which their similarity is being calculated, *N* is the total number of concepts extracted from Wikipedia (eg. present in the term to concept map) and $|C1|$ & $|C2|$ represents the number of links to other concepts from *C1* or *C2* and $|C1 \cap C2|$ represents the number of links to other concepts that *C1* & *C2* have in common.

$$\mathrm{Re}levanceScore(n,t) = \frac{\sum_{c \in C} SIM_{t,c}}{|C|} \times Commonness(n,t) \quad \text{Eq. 3}$$

Where $c \in C$ are the context concepts for the current document, *T* is the candidate concept which the relevance score is being calculated and *n* is the important term/n-gram that the candidate concept is related to. Further information on these equations is available in [Medelyan, 2008].

3.4 Document Index Querying & Searching

At the other end of text mining is querying by end users. In order to make use of the enhancements introduced during the indexing of the document collection it is necessary to enhance the query generation process. In our proposed approach, instead of converting a user's query into a pure concept query, we build an enhanced query which contains the initial term/phrase based query for text matching, but also contains, where available concepts identified as being relevant (using approaches to be described in this section), for matching with the concept information stored within the index. While this approach is based on the idea of query expansion, it goes beyond simply trying to identifying extra suitable terms to add to the query. Firstly, both terms and concepts will be included in a single query through the use of sub-queries and each sub-query component can be targeted against different fields within the document index.

Figure 4 shows our proposed approach for building hybrid term-concept queries that will be used to search on the document index. We have two methods for enhancing the query with concepts; from the initial query (QC) and from the initial results (RC).

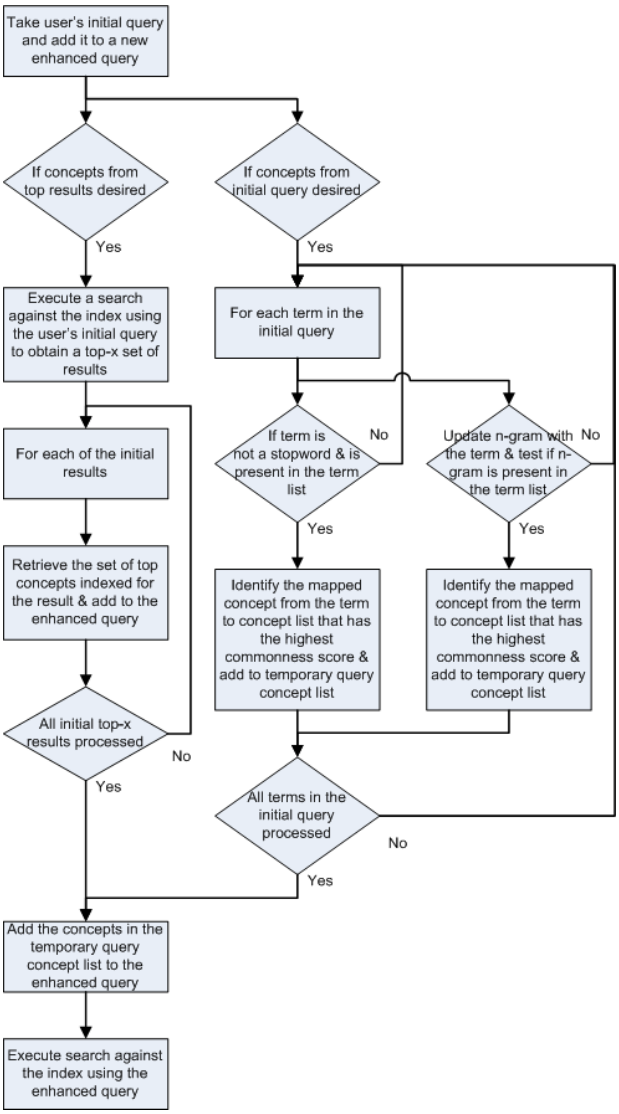

Figure 4. Overall query enhancement approach.

The first proposed method is to use the query provided by the user to obtain an initial set of search results. From these initial results we take the top-n and re-query the index to retrieve the top concepts associated with each returned result. The set of concepts from each result then forms a complete sub query of the final hybrid query. This method will work best when the initial term based query provided is of good quality and the top-n results returned from a search using this query are relevant. This helps to ensure that the top concepts utilised are relevant to the original initial query.

The second proposed method to enhancing the query with concepts is to take the initial query and in a manner similar to how we discovered and mapped concepts to a document, we discover concepts relevant to the terms and phrases present in the query. Thus for each term & n-gram we test to see whether it exists in the list of important terms that is within the concept space. If it is present, then that important term maps to at least one concept. If it maps to more than one concept we select the concept with the highest commonness score and these concepts are brought together to form a complete sub query of the final hybrid query.

The commonness score is chosen as the selector of which concept is relevant to the query because it would

be very unlikely that an accurate relevance score between a query and a concept could be calculated as there is a high probability that there would not be a suitable set of context concepts. The context concept set is required in order to be able to measure relevance if we treat the query like a document. The commonness score represents what proportion of the time; a given term or n-gram is related to a particular concept within Wikipedia.

These two approaches to creating a hybrid query can be used separately or together as each generates one or more sub components for the final enhanced query and do not depend on each other in any form. Thus there are three different combinations; 1) from the initial query, 2) from the initial results and 3) from both the initial query & initial results; that our hybrid query can obtain its concepts from. In the experiments that follow, we undertake tests on all three concept source configurations to get a measure of how well they perform in obtaining and adding relevant concepts to the hybrid query to improve search result relevance.

4 Experiments

In this section we describe the experiments undertaken to test the performance of our proposed document collection and query enhancement approach. We provide a brief summary of the real life industry dataset used along with key information, the evaluation measures used to assess performance and finally the actual results of our experiments.

4.1 Information on Datasets

The dataset used in this case study and experiments was provided by the project's industry partner, which we labelled IPDC randomly.

This dataset was built by the project's industry partner using an automated system to take in a list of starting website addresses and collect a set of pages reached via hyperlinks from the starting point. All together there were approximately 120 starting URL's supplied to the web crawler. At the end of the document collection process using these starting URL's a total of 467,070 documents had been gathered. The majority of these documents are web pages, but the collection also included pdf and word files.

Before we generated any indexes of this set of documents we undertook some basic pre-processing to identify and remove as much web page mark-up as possible. This included the identification and removal of contents such as HTML tags and Java script.

For the IPDC set, we created a small set of finance oriented queries which were used against this set. Due to the nature of this document collection we had no relevance information to draw upon to judge the performance of said queries. Thus it was necessary to manually review the results and make relevance judgments. Due to limitations in time and resources, we could only perform this manual judgment for a small number of queries and only on the results returned.

4.2 Evaluation Measures

Ideally, in such experiments we would measure at a minimum the precision and recall performance of each of the queries at various ranks. From these measurements we would then expand to other measures such as average precision and mean average precision (MAP) for each query and then an overall value for the set of queries.

For the IPDC dataset, we are only able to calculate the precision of each query and an overall average precision for the set of queries. We are unable to determine the recall result of our queries as we did not have the resources to thoroughly assess the dataset to find 'all' relevant documents to a given query. Further, the industry partner did not have the resources to undertake activity to build such a baseline to identify documents relevant to a base set of queries (in the manner of TREC for example).

We also calculated the overall average result overlap between the baseline and each experimental enhancement configuration, allowing us to discover how many results each enhanced approach had in common with the un-enhanced approach. Along with the overlap, we also measured the Kendall correlation between each enhanced approach and the baseline to gain an idea of whether there was strong, or any consistency in the actual ranking order of the documents returned in response to queries.

For the precision calculations, we manually assessed the top-20 documents returned for each query in each experimental configuration. This limit was chosen due to the cost in resources for performing manual assessment and that the top-20 results often correspond to the first page from search engines. For the overlap and correlation calculations we measured this using the top-100 documents returned for each query in each experimental configuration.

Our baseline query approach that we compare our hybrid queries against is the straight forward, simple term based query. The query terms that form each complete query in the baseline, also serve as the initial query from which the hybrid queries are built. All baseline queries search against the document's textual content and do not take advantage of any extra knowledge available by having concepts mapped to the documents.

4.3 Experimental Results

For the results in the following tables the following descriptions apply, indicating what data within the document index is being searched against for the concept component of the hybrid query;

- QC – query concepts, where concepts are discovered from the initial query and used to build the hybrid query
- RC – result concepts, where the initial query is executed and the top-x results have their top concepts extracted and used to build the hybrid query
- T3 – top-3, the top-n number of initial resulting documents whose top concepts are extracted to build the hybrid query
- B – body, the document's textual contents
- FC – full concepts, the document's complete set of concepts
- TC – top concepts, the document's set of top 10 concepts

Table 1. Overall average precision and '%' difference against baseline.

Query Exp. Config.	Overall Average Precision @ Top-20	% Difference with Baseline
Baseline	0.583	
QC_B	0.7	20.07
QC_FC	0.467	-19.9
QC_TC	0.533	-8.58
RC_T3_B	0.5	-14.24
RC_T3_FC	0.667	14.41
RC_T3_TC	**0.783**	**34.31**
RC_T3_QC_B	0.717	22.98
RC_T3_QC_FC	0.65	11.49
RC_T3_QC_TC	0.767	31.56

Table 1shows the results of our experimentation on the IPDC corpus where we executed a set of queries and manually judged the relevancy of the results returned, due to the lack of known ground truth for the corpus. The baseline query configuration; which is term based; achieved an overall average precision of 0.583 at the top-20 rank. Measured against this, were 9 experimental approaches which obtained/discovered concepts and utilized said concept knowledge in different ways.

The poor performance of two of the QC based approaches; where the concepts used in the hybrid query come from the initial query itself; comes down to two possibilities, or a combination of. First that there could have been a mismatch between the concepts selected at query time (via the commonness score) and those selected at index time (via relevance score). The second possibility is that our approach was unable to find mappings to concepts for the terms in the initial query; eg. they were not 'important terms'.

The two configurations with the greatest improvement; of over 30%; both utilise concepts extracted from an initial set of results discovered via the initial term query and then utilise said concepts against the top concepts field in the index. Thus when the initial query returns relevant results, there is a high probability that relevant concepts can be extracted from these results and added to the hybrid query, yielding further relevant results. Further, the best performance is obtained when we take the concept component of our hybrid query and use it to search against the indexed top concepts field, rather than the indexed full concepts field. This demonstrates the need for ensuring that the concepts associated with documents are strongly relevant to the document and that keeping a full list of concepts introduces weakly relevant concepts that have the potential to limit the quality of results. Further, it also demonstrates that the use of frequency; how many times a concept is mapped to a document through its various important terms; is a viable method for determining the most relevant, and hence the top concepts for a document.

We also tested using the concepts in the hybrid query and searching against the textual contents of the documents; eg. like we do with terms. The results do show an improvement is possible, indicating that in some cases the concepts themselves are present within the document's textual content; body; as terms and thus the hybrid query acts more like a term expanded query, rather than a term-concept query. However, a decrease in performance is also possible if the concepts are not present in the document's content then their similarity score would tend to decrease. Further, should the concepts be present in documents that are not relevant and do not feature the initial terms, they may be promoted up the result rankings allowing them to contend for being included in the top-20.

Table 2. Overall average result overlap and Kendall correlation against baseline.

Query Exp. Config.	Overall Average Overlap with Baseline @Top-100	Overall Average Kendall Correlation with Baseline @Top-100
QC_B	90.67	0.609
QC_FC	32.67	-0.103
QC_TC	41	-0.014
RC_T3_B	40.67	-0.083
RC_T3_FC	25.33	-0.069
RC_T3_TC	26.33	-0.079
RC_T3_QC_B	34.33	-0.135
RC_T3_QC_FC	**17**	-0.059
RC_T3_QC_TC	18.67	0.008

As can be seen in Table 2 the different methods obtaining concepts for inclusion into the hybrid query and their utilisation have differing extents of overlap with the baseline method; straightforward term based querying. The overlap ranges from a high of 90.67 to a low of 17 results in common. The fact that the overlap is not 100% makes it clear that the hybrid queries are making the determination that other documents; that the baseline discards; are relevant and should be returned. Thus the hybrid queries are not simply reordering the top-100 results. The implication of this is that a term based query here finds one set of documents, while the hybrid query finds a different set from the document collection, offering an approach for a user to find potentially relevant results that their original query would not find.

Result overlap with the baseline set is minimised when the concepts in the hybrid query are utilised against the full concepts of the documents. This happens as this field maximises the probability of matching the query concepts to document concepts; as all concepts deemed relevant to the document are present. The overlap when using the top concepts field is generally only marginally higher, indicating that even the shorter list of document concepts can be enough to allow the hybrid query to have a high probability of matching concepts. Thus both of these offer a good method for finding alternative results. The experimental approach in Table 1 that had the highest overall average precision; RC_T3_TC; only has

approximately a quarter of its results in common with the baseline. Thus not only is it ranking a larger number of relevant documents in the top-20; it is also finding documents that the baseline does not discover, opening up new sources to the user. The experiments in which the document's textual content is used for concept searching have the highest result overlap. Again, this can come down to two possibilities; first that the concepts are present in the textual content of the same documents that the terms appear in (thus reinforcing their perceived relevance) or second, that the concepts cannot be found in the textual content of most documents in the collection and the hybrid query essentially defaults to a term based query, identical to that used in the baseline.

The correlation scores also show the difference between the baseline and enhanced methods. With the exception of one enhanced method, all of the correlation scores are closer to zero, indicating a degree of independence between the baseline and enhanced methods. This indicates that the ranking order of the results from the hybrid query is different and that the concept component plays an important role in the ranking order. The exception to this is the hybrid query configuration where the concepts are obtained from the initial query terms and searched for against the document's textual content; body. Here a high correlation score showing strong agreement was obtained. This would be caused by this particular approach having difficulties in obtaining concepts from the initial terms; eg. a lack of important terms; and then locating them within the body of a document; either they are present in the same documents as the terms, or are not present at all in the majority of the documents in the collection.

5 Conclusions

In this paper we have applied the idea of using concepts, derived from a general source to build a topic specific concept space and a hybrid based querying method to then search upon an associated document index to take advantage of any topic specific concepts identified as relevant to documents within. We then applied our proposed approach to a real world industry document collection that the project's industry partner built to assist their business operations.

In our experiments we tested three methods for identifying concepts to add to the query and then tested these combinations by using the query concepts against three different fields (content, full concepts and top concepts) within the index. This helped us to discover which method of obtaining concepts and how we should use them for searching on the index. We achieved a 34% improvement in result relevancy through the use of extracting concepts from the top-3 results returned by the user's initial term-based query and using those concepts to search against the top-10 concepts associated with the documents.

Our application of the proposed methods and experiments demonstrate that there is potential for a topic specific concept space to be built from a general knowledge source like Wikipedia and that hybrid queries, composed of terms & concepts can return more relevant results than just term based query. We also demonstrate that this method can work on industry document collections via our experiments which focused on a real world collection.

6 Acknowledgement

The content presented in this paper is part of an ongoing cooperative study between Queensland University of Technology and an Industry Partner, with sponsorship from the Cooperative Research Centre for Smart Services (CRC-SS). The authors wish to acknowledge CRC-SS for funding this work and the Industry Partner for providing data. The views presented in this paper are of the authors and not necessarily the views of the organizations.

7 References

Carpineto, C. & Romano, G. (2012): A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.,* 44**,** 1-50.

Egozi, O., Markovitch, S. & Gabrilovich, E. (2011): Concept-Based Information Retrieval Using Explicit Semantic Analysis. *ACM Transactions on Information Systems (TOIS),* 29**,** 1-34.

Fang, H., Tao, T. & Zhai, C. (2004): A formal study of information retrieval heuristics. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* Sheffield, United Kingdom: ACM.

Hou, J. & Nayak, R. (2013): A concept-based retrieval method for entity-oriented search. *Proceedings of the 11th Australasian Data Mining Conference (AusDM 2013)*. Canberra, Australia.

Huang, A., Milne, D., Frank, E. & Witten, I. (2009): Clustering Documents Using a Wikipedia-Based Concept Representation. *In:* Theeramunkong, T., Kijsirikul, B., Cercone, N. & Ho, T.-B. (eds.) *Advances in Knowledge Discovery and Data Mining.* Springer Berlin Heidelberg.

Lv, Y. & Zhai, C. (2011): Lower-bounding term frequency normalization. *Proceedings of the 20th ACM international conference on Information and knowledge management.* Glasgow, Scotland, UK: ACM.

Markó, K., Hahn, U., Schulz, S., Daumke, P. & Nohama, P. (2004): Interlingual Indexing across Different Languages. *7th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO*. France.

Medelyan, O., Witten, I. H. & Milne, D. (2008): Topic Indexing with Wikipedia. *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*. AAAI Press.

Mihalcea, R. & Csomai, A. (2007): Wikify!: linking documents to encyclopedic knowledge. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* Lisbon, Portugal: ACM.

Milne, D. & Witten, I. H. (2008): Learning to link with wikipedia. *Proceedings of the 17th ACM conference on Information and knowledge management.* Napa Valley, California, USA: ACM.