Enhance legal retrieval applications with an automatically induced knowledge base

Ka Kan Lo, Wai Lam
Department of Systems Engineering and
Engineering Management
The Chinese University of Hong Kong
Hong Kong
kklo, wlam@se.cuhk.edu.hk

ABSTRACT

In this paper, we would address a challenge faced by the legal retrieval application communities: how to build up better queries for more accurate legal information retrieval. Comparing with other retrieval applications, legal retrieval usually involves more background contexts which are documented before the retrieval process starts. This is very different from other retrieval applications such as enterprise search or medical search where the queries are fired without a well defined background. Based on the rich background in legal retrieval, we devise strategies to merge the background information with an accurate knowledge base induced by Wikipedia - the online encyclopedia, to better refine the queries so that the queries are not only relying of term co-occurrence but also on the actual relations between the entities and concepts. The refined queries contain more accurate terms by combining the background and the request text information.

1. INTRODUCTION

The e-discovery process is gaining impact in the legal communities as the recent legal case involving huge corporations and a large number of related parties and the establishments of new laws demand more accurate retrieval of critical information to serve as evidence in a particular case. The pervasiveness of digital devices taking the role of storage of all the information available is increasing demands for better retrieval technologies to retrieve relevant documents from ever expanding document set. These new challenges have aroused interests in the legal community to explore better strategies in retrieval and discovery of information from legal corpora.

But the challenge ahead is far more difficult to be overcome comparing with even a few years ago. The proliferation of the Internet and its contents to every sector of the industries has made the assembling of the document set for a particular legal discovery case a daunting task as it is now possible to include a much wider set of document including different forms of web texts into a legal corpus. The huge diversity of medium of the texts comparing with the traditional medium for storage of information is introducing another difficulty to the retrieval process. Textual contents such as blogs, emails, newsgroups, forums, instant messaging, wikis and many others have their own established structural representation formats for delivery of contents. These additional formats pose different challenges comparing with the previous generation of information storage as they are more dynamic: new types can be added and obsolete types disappeared every day and the internal structures can be changed from time to time without the control of a central authority. Another challenge ahead is the pervasiveness of the informal language content being used in these new types of media. These factors all contribute to the increasing difficulties in retrieving relevant documents from law corpora.

Even the materials that we need to search are becoming more informal and dynamic, there is still one constant factor that remains largely unchanged when conducting search - the background context. When searchers perform searches, they are subjected to a particular set of background information and concepts. Comparing with a typical search in other area such as consumer search, the background context of a typical application of legal search is more well defined. In consumer search or some other search scenario, the search application is supposed to serve a vast amount of audience for any contextual background. Typical searchers in this scenario are not interested in the actual recall of the search as the total number of documents are not their major concerns in conducting searches, instead they are more interested in precision, or the user satisfiability of the returned results.

However, searches in legal communities and especially the legal discovery are largely different from the current popular search paradigm in which recall plays a more important role than precision. Especially in legal discovery, it is more important to search deeply in the collection to discover whatever potential evidence that may be missed by simple search paradigm.

The factors of context and search depth thus appear parallel in designing a particular strategy for retrieval of information. The richer the context, the deeper one would expect the retrieval results would be.

Taking the examples of complaints and the production requests in TREC Legal track, one can easily discover that the complaints contain a lot of background information for more accurate retrieval process. The context information may be itself quite rich to help us to get a good result.

Besides considering the additional content of web texts in a legal corpus, the web can also be a source of structural knowledge for better retrieval applications. Enriching the search process with the structural knowledge would improve the search processes as the web contains a higher redundancy of information and the context of the web texts would match some of the contents in the legal corpus, which will contain an increasing portion of web texts. The contributions of this work are as follows:

- New method is proposed to automatically enrich the context
- Integration of the expanded contexts for better search retrieval procedures
- Algorithm and model to induce and build up the background knowledge base from web texts.

This paper is organized as follows: Section 2 gives the background work. Section 3 presents the investigation of the practices in delivering the legal retrieval process. Section 4 shows the algorithms in automatically inducing the knowledge base from web texts. Section 5 presents the method to expand the context information of the complaints using the external resources. Section 6 is the conclusion.

2. RELATED WORK

One stream of the related works is from the TREC legal track report [2] that documented the techniques used by the TREC Legal track participants last year. The legal track evaluation started from last year and a number of teams participated in this exciting area of retrieval research. Current strategy in formulating the queries involves the request texts where the text contents are broken down into lists of keywords, subjected to possible stemming and stopword removals, and then fed into the retrieval system.

Focusing on the ranked retrieval work in the current study, the Open Text group utilizes the request text in the ordinary production requests to formulate the queries [6]. The queries, subjected to wildcard, proximity matching are then fed to the retrieval system. The Maryland group also uses the request texts as the major text fragments as the list of keywords for the retrieval of documents [5].

Some groups use query expansion model to expand the query text for better retrieval of documents [8]. The other group utilizes the boolean query, defendents' query, plaintiffs' query and the final query as a source of request texts [7].

As it was the first year of TREC Legal track event, many groups were demonstrating their preliminary investigation of this task. However, some observations can be made in the current approaches in developing the legal retrieval applications. First, the heavy reliance on request text: Many retrieval systems depend on the simple syntactic format of the boolean, defendent queries and plaintiff queries as a source of keywords. Though initutive, however, it completely ignores the backgrounds in which the particular query is formulated. It would be beneficial if more background information can be embedded in the queries. As explained in the previous section, the search operations in the legal applications are rich in context and background information, it is possible to further explore the possibilities of using this background information to build up the queries. Second, the query expansion techniques are currently under-utilized in the retrieval approach. While it is understood that the query expansion strategies can only usually improve the accuracy of the top retrieved results based on the existing approaches, given the more precise background information, the situation may be different. Third, no attempt has been made to integrate the external source texts for retrieval. In other words, the current retrieval approach assumes that the corpus contains enough information for retrieval. As explained in the previous section, increasing content diversity from the web source expands the scope of the text information involved and it is reasonable to include this information to formulate better queries.

3. PRACTICE IN RETRIEVAL PROCESS

As mentioned in previous section, the major difference between the legal retrieval applications and the general search applications is the rich context in formulating the queries and it is hypothesized that, in this paper, this rich context can be very important in achieving a better quality of legal retrieval.

The TREC Legal track, organized by TREC last year, provided examples of how the legal retrieval is taken place. The retrieval process is taken by two parties where both the defendants and plaintiffs devise their version of queries based on the Boolean query format. The defendant proposes the queries first, the plaintiff then adds the extra terms to the queries in order to extract a larger proportion of the relevant documents from the corpus. The final queries are then fed into the system for final retrieval. The followings are the examples extracted from the production requests of the Legal track:

This case is about extracting all relevant documents about "Enchinoderm Cigarettes" companies and all the relevant documents about its placement of advertising materials in different media. A number of request texts are formulated in the current proposal and some of them are listed below:

Request Text 1
All documents discussing, referencing, or relating to company guidelines or internal approval for placement of tobacco products, logos, or signage, in television programs (network or cable), where the documents expressly refer to the programs being watched by children. (Note: "children" refers to persons under the age of 18.)

Defendant's boolean query 1
Defs.' Proposal: (guidelines OR strategies OR "internal approval") AND placement AND (logos OR signage) AND (television OR cable) AND "watched by children"

Plaintiff's boolean query 1
Pls.' Counterproposal: (guide! OR strateg! OR approval)

AND (place! OR promot! OR logos OR sign! OR merchandise) AND (TV OR "T.V." OR televis! OR cable OR network) AND (watch! OR view! W/5 (child! OR teen! OR juvenile OR kid! OR adolescent!))

Request Text 2
All documents discussing, referencing, or relating to company guidelines, strategies, or internal approval for placement of tobacco products in movies that are mentioned as G-rated.

Defendant's boolean query 2
Defs.' Proposal: (guidelines OR strategies OR "internal approval") AND placement AND "G-rated movie"

Plaintiff's boolean query 2
Pls.' Counterproposal: ((guide! OR strateg! OR approv!) AND (place! or promot!)) AND (("G-rated" OR "G rated" OR family) W/5 (movie! OR film! OR picture!))

Request Text 3
All documents discussing, referencing or relating to company guidelines, strategies, or internal approval for placement of tobacco products in live theater productions.

Defendant's boolean query 3
Defs.' Proposal: (guidelines OR strategies OR "internal approval") AND placement AND ("live theater" OR "live theatre")

Plaintiff's boolean query 3
Pls.' Counterproposal: ((guide! OR strateg! OR approv!) AND (place! or promot!) AND (live W/5 (theatre OR theater OR audience"))

Request Text 4
All documents discussing, referencing, or relating to payment or compensation to 20th Century Fox Corporation for placement of products and/or brands in a film production. Compensation should be interpreted as monetary payment, goods, services, or other considerations.

Defendant's boolean query 4
Defs.' Proposal: (payment OR consideration) AND placement AND "20th Century Fox Corp!" AND "film production"

Plaintiff's boolean query 4
Pls.' Counterproposal: (pay! OR paid OR compensate! OR consideration) AND ("20th Century Fox" OR Fox OR Newscorp) AND (film! or movie! or production)

Request Text 5
All documents discussing, referencing, or relating to budgets, actual costs, or planned costs for placement of products and/or brands in either television or film media, which expressly reference or discuss yearly expenditures for product placement.

Defendant's boolean query 5
Defs.' Proposal: (budgets OR "actual costs" or "planned costs") AND placement AND (television OR films) AND "yearly expenditures"

Plaintiff's boolean query 5
Pls.' Counterproposal: (((budget w/5 (actual OR plan!)) OR costs) AND place! AND (TV or "T.V." OR "televis! OR cable or network) AND "((yearly OR annual) W/5 expen!)"

However, current proposal in retrieval does not take into account the background of the retrieval process and merging of these backgrounds with the more general web context to deliver a better query. Consider the first complaint, the background, which is reproduced here, provides far more information then the request texts and the queries alone.
"According to information and belief, Echindoernm Cigarettes and other companies have a long history of placement of tobacco products and brand images in the public media. These media, including television (network cable), film, a live theater, and rock concerts, are regularly viewed by children, teenagers, and young adults. Such individuals are at the most impressionable time of their lives, and are unknowingly exposed to de facto advertising for tobacco and tobacco-related products simply by watching such media.
In particular, the glamorous manner in which smoking and other tobacco use are portrayed on the screen adds a cachet to the habit that encourages young people to try smoking for the first time. Thus is exposed the true motivation for product placement - inducing non-smokers to become smokers with blatant disregard for the long term effects and public health risks associated with tobacco use."

4. GENERATION OF BACKGROUND CONCEPTS

To enrich the context of the queries, we rely on some high-level knowledge sources to automatically improve the retrieval accuracy. The knowledge source currently used are from Wikipedia in our experiments. However, we expect that more different sources such as high-level ontologies would be used in the future.

4.1 Formalization

The structure of the knowledge representation is formalized in this section. Comparing with other approaches, relative clean structures are used in this work.
The basic structures used are the entity and relation, unlike other works using which use the similar kind of primitives, no further classification schemes are currently used under the entity and relation class. These rather unconventional structures are proposed to faciltate the procedures in extracting knowledge from texts and in particular, the linked style texts in an unsupervised fashion. In some other fine-grain classification scheme, the information under interests are generally in five main classes: entity, relation, event, temporal expression and value where further classes, subclasses and types are defined under these five main classes. In other hand-crafted knowledge source, it is quite common to have more finer-grain ontologies to describe the relations between the concepts. Figure 1 shows an example of the fine-grain structure as in [3, 1], which is the definition of "action" class where detailed structures have been proposed

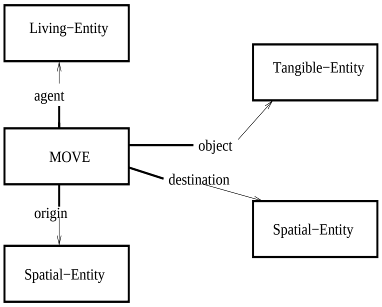

Figure 1: Example of a fine-grain structure which contains detailed information about the relationships with other entities

to relate this and other classes such as entity.

However, defining such a fine-grain structure can introduce difficulties in constructing the contents of these structures automatically, and in particular from texts. The reason why fine-grain structures are proposed is that is that in the general knowledge system where the machines do not have any knowledge, knowledge engineers will not only encode the concepts and their lexical realization but also the relations between different concepts. These relations are encoded in the “arguments”, “slots” of a particular class. However, in syntactic representation of texts, these relations are generally implicit in the grammatical relations. The definition of the grammatical relations are generally used for defining specific roles in the syntactic form, not in the semantic relations and thus inconsistencies exist between the definitions of relations from both sides. Another problem in using such a fine-grain structure is that the fine-grain relations seldom exist in a direct correspondance with the grammatical relations. Thus, simple structures are preferred in this work.

4.1.1 Text Structures

The encyclopedia texts can be considered as a huge graph connecting different concepts together, with the graph nodes denoting the concepts and the links between the pages denoting the relationships between concepts. These concepts may be proper nouns, common nouns, entities, relations, verbs and so on, depending whatever the contributors can think of. Under each concept, description text is written to describe the concept. In the description texts, “links” are made to connect the current concept or the current context in the particular fragment of texts to the other concepts. These “links” are not generated automatically but explicitly made by the contributors who write the particular piece of texts. In other words, the contributors take a similar role as the knowledge engineers. Instead of using specialize version of coding scheme, they use the texts and links to code the knowledge. Figure 2 shows a fragment of the structures of the encyclopedia plotted. Figure 3 shows some of the statistics of the structures of the encyclopedia text as in September, 2006.

Definition 1: Encyclopedia Structure
$page_i$: The text entries in the encyclopedia,
$title_i$: The title string of a page,
$sentence_i$: Sentences within a page with words w_i

$E = \{page_i\}$
W: The set of words in encyclopedia
$TITLE$: The set of titles of encyclopedia pages
$page_i =< title_i, \{sentence_{ij}\}, \{page_{ik}\} >$
$sentence_i =<< w_1, w_2, ..., w_n >, page_{ik} >$
$title_i =< w_1, w_2, ... >$
where $w_i, w \in W$ and
$page_{ik} \rightarrow page_i, page_i \in E$

This definition states that the text structure within the encyclopedia consists of pages linking together with each page representing a node in the graph.

Number of Articles	1.4 million
Number of words	609 million
Mean article length	435 words
Number of characters	3.5 billion
External links to other websites	2.6 million
Cross reference links	32.1 million
Number of contributors	151934

Figure 3: Statistics of the English Wikipedia in September, 2006

4.1.2 Modeling Primitives

In designing the modeling primitives in the framework, simple structures are used. As the structures themselves have a close connection to the underlying text, constructs are defined to faciltate the lexical realization of the concepts from texts. Two classes of entites and relations are defined.
Definition 2: Concepts
The concept, c, denoted by the word $w \in W$, is represented by:
$w \rightarrow c$ and

Definition 3: Surface Concepts
A surface concept is defined by:
$\forall w$ where $w \rightarrow c$
$w \in TITLE$
This means that the surface concepts consist of the set of title words in the pages of encyclopedia.

Definition 4: Hidden Concepts
A hidden concept is defined by:
$\forall w$ where $w \rightarrow c$
$w \notin TITLE$

4.1.3 Entities

The class of entities represents the most basic objects described in the texts. These objects include name entities, locations, organizations and so on. They are those objects participating in the roles of the relations.

Definition 5: Entity
An entity e is defined by:
$e \subset c$
and $e \neq r$, $e \in role$ of r

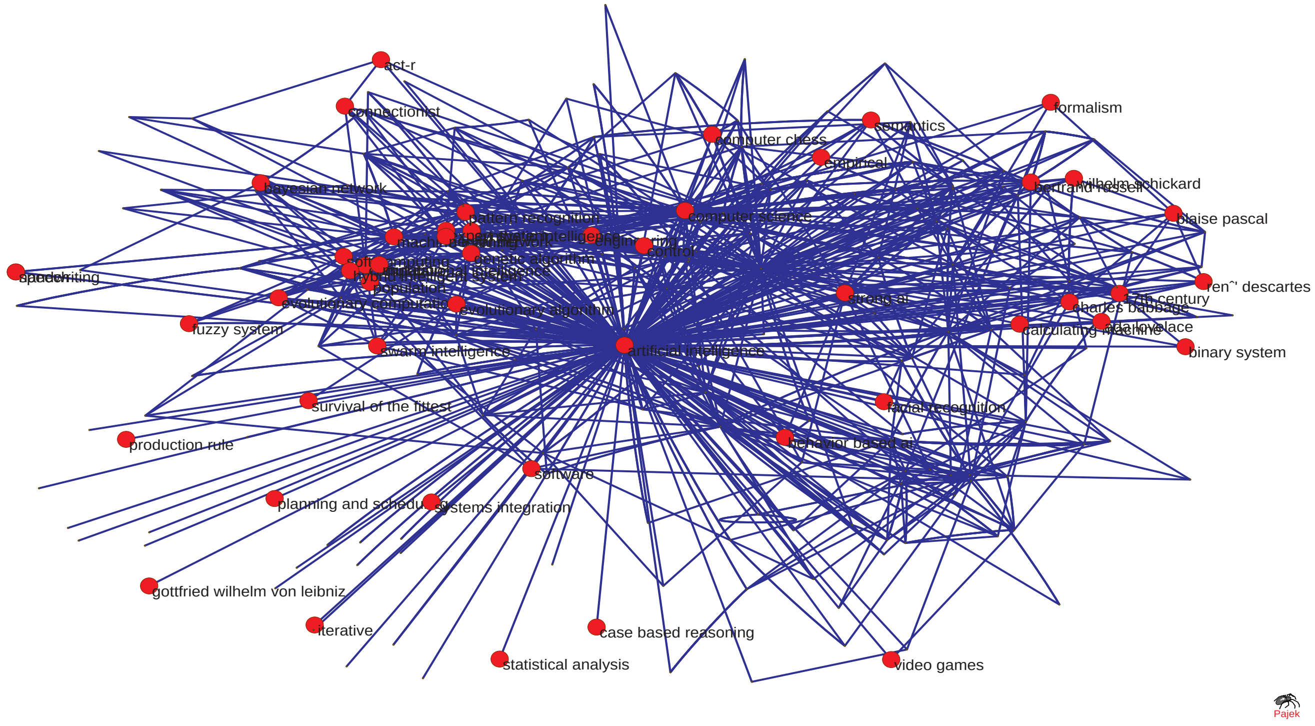

Figure 2: A graph showing the encyclopedia text structure with 200 nodes extracted in the domain of "Artificial Intelligence"

Entities examples	Relations examples
Indianapolis Colts	jump
Super Bowl XLI	run
global warming	drink
Lumber Exchange Building	offer
City of Sawson Creek	sleep
Annadel State Part	stop
Fort Canning Tunnel	discover
Tesla Coil	depend
Peace River Regional District	check
Antaectica	

Figure 4: Entities and relations examples extracted from the encyclopedia

where r denotes the relations and roles denote the arguments of the relation r

Figure 4 shows some of the entities in the encyclopedia text.

4.1.4 Relations

The class of relation represents the relationships between concepts or terms. Instead of using labeled relations, the relations are unlabeled and their existences are represented by the entites and the respective position of roles they participate in.

Definition 6: Relations
A relation r is defined by:
$r \subset c$
and $r \neq e$,
$r =< role_1, role_2, role_3 >$
where $role_i = e$ and e is the entity

Figure 4 shows some of the relations in the encyclopedia text.

4.2 Semantical Domain

Comparing with ordinary texts, the structures of the encyclopedia not only provides relationships between surface entities and surface relations. The link structure also describes under which area of the graph or domain a particular concept realizes. This is depicted in the graph as shown in figure 2 containing the concepts such as "Artificial Intelligence", "Knowledge" and others.

Given a particular list of concepts, a list of other concepts can be extracted based on the "link" between concepts. Depending on the concepts chosen, the linkages between the concepts in the texts, the "knowledge" and the word usages by the contributors, different text segments can be combined and merged under different sets of concepts. These particular fragments enable a large number of specialize domains to be generated where the general domain can be built by joining all these specialize domains together.

Definition 7: Semantical Domains and their Operations
A semantical domain D is defined by:
$D = \{page_i\}$ where $page_i \in E$ and

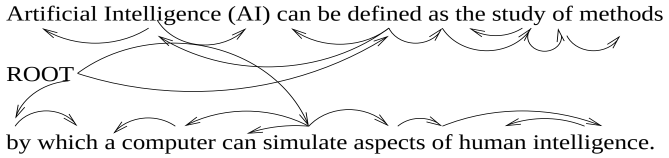
Artificial Intelligence (AI) can be defined as the study of methods
ROOT
by which a computer can simulate aspects of human intelligence.

Figure 5: Dependency parsing of sentence

$\forall page_i \rightarrow page_{ik}$ and $page_{ik} \in D$

A semantical domain thus consists of a group of pages that forms a connected graph with the links. From the perspective of the texts, D consists of a set of texts where their meanings are related to the others.
Some definitions of the operations on D are defined:

Two domains: D_m and D_n overlap if:
$\exists page_i \in D_m$ and $page_i \in D_n$

Domain D_m is subsumed by domain D_n, denotes by $D_m \subset D_n$ if:
$\forall page_i \in D_m, page_i \in D_n$ and $D_m \neq D_n$

Domain D_k is formed by joining domain D_m and D_n where:
$D_k = \bigcup(page_i \in D_m, page_j \in D_n)$
and denotes by $D_k = D_m \oplus D_n$

Based on the internal structures of entities, relations and the operations of domains, their respective contents are filled from the texts.

4.3 Knowledge Extraction

As the knowledge is encoded in the text, the respective concepts have to be extracted from the texts to fill into the entity and relation objects. In this work, we focus on extracting surface entities and surface relations as more hints are provided for these kinds of concepts with the explicit linkages.
Instead of using linguistically rich syntactic representation, we use simpler syntactic representation such as the unlabeled dependency grammatical structure to extract the syntactic structure from texts [4]. The major reason in using simpler structures is that there are a lot of extra information residing in the linguistically rich structure that are generally irrelevant to the construction of the entities and relations. Figure 5 shows the parse result of a sentence where a more direct correspondence of texts with the entities and relations can be found.

4.4 Syntactic parsing of sentences

To extract the relevant items, the texts are first parsed using dependency syntactic parser. Dependency structures are produced for each sentence under a particular domain.

Definition 8: Parse Representation of Sentences
A parse of the sentence with words w_i is defined as: $sentence = \{w_i : (i \mapsto j) \in L\}$
where w_i is the word at position i with final parse of the sentence denoted by L. In this definition, w_i is the head while w_j is the tail.

Instead of using the labeled representation from the parser such as part-of-speech, grammatical relations and so on. Only the head-dependent relations and the relative position of the roles are used in the extraction process. This reduces the burden of decoding a rich representation which may introduce extra error in the further extraction process.

4.5 Entities and Relations

After the parsing process, large number of sentences are collected. These sentences contain large number of entities, relations and so on. For example, after processing the texts under the concepts of “artificial intelligence” and “knowledge”, a number of potential entities and relations appear in the text as depicted in figure 2.

Using these sentences, the algorithms as shown in figure 6 are run to build up the entities and their relations from texts. The algorithm works by extracting the head and dependent relations from the parsed sentences and makes hypothesis that linked entities are in a relations. By accumulating the instance of entities and relations from a large amount of texts in a particular semantical domain, the interrelationships between these items are gradually built up.

Algorithm: (Extraction Process)

```
For $\forall sentence_{ij}$ with parse $L$ where
    If $\exists w_k \in sentence_{ij}$ and $w_k \in TITLE$
        For $w_k : w_x \rightarrow w_k$ where $w_x \in sentence_{ij}$ and
        $w_k \mapsto \emptyset$
            Build $e$ where $w_k \rightarrow e$ and
        For $w_k : w_k \rightarrow w_x$ where $w_x \in sentence_{ij}$
        and $w_1 << w_2 << w_3$
        where $w_i << w_j$ means that $w_i$ precedes $w_j$ and
            $w_k \mapsto w_1$,
            $w_k \mapsto w_2$ and
            $w_k \mapsto w_3$

        Build $r$ where
            $role_1 = e_1$ where $w_1 \rightarrow e_1$
            $role_2 = e_2$ where $w_2 \rightarrow e_2$
            $role_3 = e_3$ where $w_3 \rightarrow e_3$
    End if
```

Figure 6: Algorithm description of the extraction process of surface concept

Based on this method, a large number of entities and relations are extracted as shown in the example in figure 2.

5. COMBINING THE CONCEPTS WITH THE BACKGROUND CONTEXT FOR BETTER LEGAL RETRIEVAL EXPERIENCE

The induced entities and relations are of high quality as there have been a lot of conceptual linkages built by people instead of using classification by machine learning algorithms in the type of texts we use for knowledge induction. Thus, these entities and relations represent a high level of relationships between different concepts. This is especially critical in legal retrieval process where the document size is large, ranged from millions to even billions of documents, employing pure keyword-based query, without the necessary backgrounds to direct the retrieval process, would only retrieve a large list of documents without precisely attacking the recall problems as focused in legal retrieval. Projecting the background information and request texts, back to this large network of relations, and then refining the query to contain this high level of concepts, would have the potential in obtaining better results.

5.1 Algorithm

The following algorithm describes the process of enriching the background context with the knowledge base generated as described in the previous section.

- From the background text and request text, remove the stopword and perform stemming
- Match each of the remaining terms in the background and request text to the Wikipedia text articles, indexed by the title of the articles.
- Each of the expansion of the titles would represent a network of concepts and relations, as described in the previous sections.
- From the networks extracted, combine each of these networks to a unified network and filter those terms that do not connect to any of the concepts in the unified network.
- Extracting the terms and concepts which are of high accuracy to the query strings
- Fire the query to the retrieval system

6. CONCLUSION

This paper addresses the challenge of legal retrieval process by first explaining the unique challenges faced by the legal retrieval applications: the diverse source of texts, huge document collections and the difficulties in formulating the precise queries for retrieval. Our solution is to project the background information and the request texts, which form the background context of the construction of queries, to a network of concepts, including entities and relations, which are of high qualities, and then to refine the queries so that it can embed more concepts related to the background information and the request texts, to improve the performance of accuracy.

We are testing the performance of this system, using the materials of the TREC 2006 Legal track evaluation. The future work may include refining the filtering process of the merging algorithm so that the terms selected would be of higher qualities and thus, may lead to more accurate retrieval performance.

7. REFERENCES

[1] K. Barker, B. Porter, and P. Clark. A library of generic concepts for composing knowledge bases. In *Proc. 1st Int Conf on Knowledge Capture (K-Cap'01), 2001.*, 2001.

[2] J. Baron, D. Lewis, and D. Oard. Trec 2006 legal track overview. In *The fifteenth Text Retrieval Conference (TREC2006) Proceedings*, 2006.

[3] P. Clark and B. W. Porter. Building concept representations from reusable components. In *AAAI/IAAI*, pages 369–376, 1997.

[4] R. A. Hudson. *English Word Grammar*. B. Blackwell, Oxford, UK, 1990.

[5] D. Oard, T. Elsayed, J. Wang, Y. Wu, P. Zhang, E. Abels, J. Lin, and D. Soergel. Trec 2006 at maryland: Blog, enterprise, legal and qa tracks. In *The fifteenth Text Retrieval Conference (TREC2006) Proceedings*, 2006.

[6] S. Tomlinson. Experiments with the negotiated boolean queries of the trec 2006 legal discovery track. In *The fifteenth Text Retrieval Conference (TREC2006) Proceedings*, 2006.

[7] M. Wen and X. Huang. York university at trec 2006: Legal track. In *The fifteenth Text Retrieval Conference (TREC2006) Proceedings*, 2006.

[8] F. C. Zhao, Y. Lee, and D. Medhi. Experiments with query expansion at trec 2006 legal track. In *The fifteenth Text Retrieval Conference (TREC2006) Proceedings*, 2006.