Abstract
Legal professionals around the world often reference and use online document libraries such as WorldLII to conduct legal research. In particular, under the legal system in Hong Kong, relevant judicial precedents are taken into consideration during judgements, highlighting the importance of retrieving historical judgements during legal research. This project will study, implement and evaluate various algorithms and techniques to introducing semantic querying and document summarization to the Hong Kong Legal Information Institute (HKLII), with aims to facilitate the searching and reading of documents on the system. Based on previous research work, this project will focus on using topic modelling techniques, while also potentially exploring other algorithms, to achieve document summarization by providing information that is semantically the most relevant to the user. Different categories of judgements will attain different types of relevant information, and such data is identified based on domain knowledge from legal experts in Hong Kong. By achieving document summarization and semantic searching on HKLII, it is hoped that the system can better understand user queries and retrieve more relevant documents, while also reducing the time needed to recognize key information of the judgement, allowing users to more efficiently identify the relevance of the judgement to their needs.
1 Introduction

In common law, there is a great emphasis on resolving disputes with reference to previous court decisions. As a region that practices common law, Hong Kong has a legal structure that makes court decisions legally binding on hierarchically inferior courts. Thus, researching historically relevant judgements plays an important role to the legal community, and the Hong Kong Legal Information Institute (HKLII) is a source of research materials for legal professionals.

1.1 Current status of HKLII

HKLII is a document database that provides the general public with free access to primary legal materials. Its deployment and infrastructure, including the search engine and database structure, is provided by Australasian Legal Information Institute (AustLII), the Australasian counterpart of the worldwide Legal Information Institute systems. Currently, HKLII supports searching of legal judgements for retrieval, but is limited to simple lexical searches. Visit logs of HKLII have revealed that about three-quarters of visitors to the website made queries using the search bar or advanced search functions. Studying the specific search logs, one can find that among the queries, case search, entity search and concept search each account for about a quarter of the queries. Case search and entity search refer to queries related to a particular historical case and person or organization respectively. In this scenario, it is not difficult to find the judgements users want in a keyword-based search because of the uniqueness of terms such as names and case identifiers in these queries. On the other hand, concept searches such as Companies Ordinance or Inland Revenue Ordinance concern the judgements relevant or exemplary of legal concepts on a higher level, which are more difficult to be captured accurately by a non-expressive lexical search. Furthermore, based on activity logs of the website, it is shown that concept searches take the longest average time before users arrive at a desired document. Apart from the poor ranking in keyword search results, which makes it more likely to require multiple searches before retrieving the desired article, lengthy search times can also be attributed to users requiring more time to identify if the judgement suits their needs, unlike case searches or entity searches where users can easily tell the difference. Introducing semantic search improves the ranking of the judgements, while document summarization will shorten the time needed to understand the article and determine the relevance to users’ needs, overall improving the user experience in searching legal documents.

1.2 Semantic Search

When querying an abstract legal concept, the user might not be able to provide the exact wordings or technical terms of the concept, resulting in the keyword search engine missing out on the judgements that the user might want. For example, wounded hand only returns 2 results, while hand injury and injured hand returns 66 and 39 results respectively. It is noteworthy that some results in injured hand are not present in hand injury queries, and vice versa, so potentially relevant results are not brought up to the user if the user does either of the searches, and such situations can be best avoided with the aid of semantic searches. Semantic searches can be achieved using different algorithms, and the focus of this project will be on using various topic modelling techniques to identify clusters of keywords that have strong associations by analyzing documents in HKLII and turn these keywords to information.
in semantic search. For example, one can associate words of *finger injury* by its synonyms or more specific terms of fingers, such as hand, thumb and joint. Such clusters can be classified as conceptual topics (in this case, a topic that describes words related to a finger injury). The plan of the project is to relate judgements to certain topics, and when the query arises, the most relevant topics to the query will be identified and matched against the topics in the judgements to rank the most relevant judgements to the user. In theory, the more coverage a topic has in a judgement, the more relevant the judgement should be in the context of the topic.

1.3 Document Summarization

In the current HKLII interface, users will only be able to see the title of the judgements before clicking in the hyperlink text to inspect the document in detail. Such design is not very telling because users are unable to see why the search engine included this result. In contrast, popular search engines like Google often highlight words in the description section of each query result, revealing some detail of the search result and helping the user determine the usefulness of the document before viewing in full. This is similar to creating *query-driven summarizations* in this project, as this function aims to save users from having to read through the entire case, only to realize that the case does not suit their intents. Query-driven summarizations highlight the most relevant parts of the judgement related to the topics in the user’s query to justify why the judgements have been returned. Once the user has found the case useful, *aspect-driven summarization* comes in to help the user extract the most important information of the case. While the importance of different categories of cases vary, information generally perceived as important for certain categories of judgements have been collected by consulting legal experts on their opinions. Thus, the most commonly used information of a category is identified and shown to the user. For instance, in *personal injury* cases, it is usually the case that the background, treatment, losses and compensations of the plaintiff are of the most interest for legal researchers. This type of summarization aims to provide a summary of the above-mentioned extracted points specific to the aspect, so crucial particulars of the case can be understood in a short period of time. For aspect-driven summarization, the focus of this project will be on *personal injury* and *drug trafficking* cases with the aid of legal expertise in such areas, as they account for most of the cases available in HKLII. Besides, the sentencing in drug trafficking cases and the quantification of compensatory damage in personal injury cases also particularly depend on past precedents, making these two categories of cases more frequently studied.

1.4 Objective

In summary, this project aims to overall *improve the user experience* of the HKLII website, in particular during the phase of document searching, by improving the ranking system and understanding queries at a semantic level, as well as providing summaries of the document, both from a query perspective and a category perspective. The project aims to achieve the goals by mainly using *topic modelling*, a technique that was investigated and implemented in past related research, while also exploring other machine learning and natural language processing algorithms and tools, to evaluate and compare the performance of different algorithms. Whenever feasible, the implementation of other (potentially proprietary) legal document database soft-
ware such as Westlaw and LexisNexis will be discussed and compared. Finally, if time allows, the project will include a simple proof-of-concept prototype that demonstrates the functionalities of semantic search and document summarization using the techniques implemented in this project, for users to better understand the improvement compared to the current HKLII. It is aimed that this functionality will eventually be integrated and deployed to the newer HKLII system, which includes a new user interface and up-to-date infrastructure, and is being developed now by a team of developers.

1.5 Related work

This project is closely related to recent work done in applying topic modelling to semantic queries and document summarization to HKLII, which has shown promising results. Along with related work on legal text processing, it is observed that Latent Dirichlet Allocation (LDA), labelled LDA (LLDA) and hierarchical LDA are amongst the topic modelling techniques adopted for use. For machine learning algorithms, techniques such as the Convolutional Neural Network (CNN) are used to predict the likelihood that any sentence in a judgement should be included in the summary. This project aims to focus on topic modelling, but also implement and evaluate different algorithms suggested in related work with some new algorithms proposed.

Besides, two other final year undergraduates of the year 2021-22, Yuchen Liu and Huijie Pan, will also work on adding features to HKLII from a natural language processing (NLP) perspective to improve the user experience of browsing documents. Their work will include introducing taggings to the documents, which includes information such as key legal concepts and unique words in the document, as well as a document recommendation system that recommends highly related cases to the reader, based on their activity on the website.

2 Methodology and Algorithms

This section highlights the implementation of the above-mentioned features in HKLII. Topic modelling will be the primary focus of the project, so the methodology of achieving semantic searching and summarization of legal documents will primarily be discussed with topic modelling techniques, with a brief inclusion of potential algorithms that can be used in a later stage of the project.

2.1 Data Collection and Development of Prototype

Access logs and legal data of the current HKLII website are available online in the HKLII database, which can be freely accessed anytime. This makes data collection in this project straightforward. By investigating the logs, one can also identify the problems in HKLII and develop additional features in the prototype to best avoid the issues from happening again. New infrastructure will also be taken into consideration during the development of prototype, granted that there is enough time, to replicate the environment of the new HKLII design as much as possible and prove the integrability of semantic search functions and document summarization functions to the new website. There is no requirement for the language of implementation for this prototype, should the development be done, but for consistency reasons,
a Python-based web framework such as Django or Flask is preferred as it is likely that Python will be used in performing topic modelling and machine learning algorithms that power the semantic search and summary functions.

2.2 Topic Modelling

Topic modelling refers to the process of extracting abstract topics from a collection of documents. Formally, given a context of documents, performing topic modelling returns a set of topics $T = \{T_1, \ldots, T_n\}$. There are a lot of topic modelling techniques that can potentially be successful to HKLII, including LDA, LLDA, hierarchal LDA, etc. Topic Modelling is a mature field with a long history and extensive research, and depending on the algorithm’s nature, some may suit more in HKLII than others. In previous research work, LDA has been selected to perform topic modelling, which generates the set of topics based on the word distribution of each topic and the topic distribution of the document (the probability of each topic occurring in the document). Three approaches of pre-processing for LDA had been discussed and put forward with varying degrees of labelling to extract the topics of a judgement. The method with lowest intervention involves only removal of numbers and stop words, while the second approach extracts manually annotated words as labelled text for LDA to remove the unimportant details of a topic. The final approach with most human effort further categorizes the important information into aspects (such as plaintiff’s background, injury, treatment and loss for personal injury cases) and information in each aspect are fed into LDA separately. It has been shown that the labelling with categorized aspects performed the best.

Semantic search then leverages these topics generated and compare them against the judgements in terms of the topics. An example of a quantitative way to measure the relevance between a given query and a judgement is to construct probability vectors $p \in \mathbb{R}^n$ for the query and each judgement in the database, with each entry $p_i$ denoting the likelihood that the query/judgement is relevant to a topic $T_i \in \{T_1, \ldots, T_n\}$. Once all the probability vectors are computed (only the probability vector of the query should be computed in query time, probability vectors of the judgements can be pre-computed), a function $f : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ that can measure the distance between the two vectors, such as the dot product, can be used to determine and rank the relevance of each judgement against the query. While this is a proven way to convert topic modelling results to ranking documents in a query, there is an abundance of alternative methods that can achieve a similar result, and throughout the process of the project, various algorithms, processing methods and functions will be studied and tested for evaluation. Improvements of these processing methods, alongside new techniques and topic modelling algorithms, will be discussed.

In terms of document summarization, the methodology for query-driven summarization and aspect-driven summarization will be different. In current research results, query-driven summarization uses the topics from topic modelling to compute the similarity between a topic and a paragraph/sentence, and selects the sentences that are most relevant to the topics of the user’s queries. Aspect-driven summarization, on the other hand, first identifies the topics of an aspect in the judgement, before comparing the paragraphs in the judgement and the topics using similar functions and algorithms. The project will build upon the current work by both trying to alter the topic modelling functions to generate different topics and the mechanism to
the processing of the topics.

2.3 Alternative algorithms

Topic modelling will be the primary focus of the project, but other algorithms will also be explored and experimented to achieve semantic searching and document summarization, among of which include the use of Convolutional Neural Network (CNN) to predict the likelihood that a sentence should be included in a summarization. Supervised Machine Learning models in achieving document summarization can be explored, with the availability of labeled data and legal expertise for this project. For semantic search, algorithms such as TextRank can also help legal researchers and users to quickly determine if a judgement suits their needs. The project will adopt suitable current practices in academia on document summarization and semantic searching, and explore in-depth their performances in HKLII.

2.4 Evaluation

The evaluation of the performance of algorithms on document summarization and semantic search will involve both quantitative and qualitative measures. With the personnel of legal experts available for the project, they will be invited to evaluate the summaries to grade the significance of each paragraph in the summary. Metrics against the optimal summary, which is a summary conducted manually, will be compared and measured quantitatively as a benchmark. As for semantic searching, actual queries in access logs and customary queries will be designed and observed qualitatively to measure the performance of the semantic search results against the existing lexical search engine in HKLII. If necessary, quantitative measures can also be raised to measure the performances of different semantic search algorithms with feedback from legal experts, by having them score the actual relevance of a judgement given a query.

3 Schedule and Milestones

With reference to the logistics of the project, the project will be divided into three main parts - the first phase (“inception” phase) includes this project plan and the construction of the project webpage. The project webpage is visible at the department’s Final Year Project Webpage database (https://wp.cs.hku.hk/fyp21033).

Next, the second phase will include the research, implementation and evaluation of various topic modelling techniques against HKLII on semantic search and document summarization. By the end of the second phase, preliminary research findings and results on topic modelling and the applications of it in HKLII will be discussed in the preliminary report.

Finally, in the third phase, topic modelling techniques will be further investigated, and other machine learning algorithms will be proposed, studied, and tested in HKLII. A proof-of-concept prototype is planned to be implemented in this stage, which should include features from both this project and work from Yuchun Liu and Huijie Pan, the two other undergraduates working on this project this year. Depending on the availability, the design and implementation of the other software systems that legal researchers use, such as Westlaw and NexisLexis are discussed.
and compared against the new version of HKLII.

References


