Searching in Chinese: the Experience of HKLII

Alex Y.H. Fung[*], Kevin K. H. Pun[*], Philip Chung[°], Andrew Mowbray[°]
[*]*Department of Computer Science, The University of Hong Kong*
[°]*Faculty of Law, University of Technology, Sydney*

Abstract. The Hong Kong Legal Information Institute (HKLII) provides the general public, free of charge, online legal information relating to Hong Kong. A unique feature of the Hong Kong legal system is its bilingual nature (English and Chinese). Hong Kong has a bilingual court system and bilingual statutory laws of equal legal status. Thus to serve the Hong Kong community, HKLII must be able to search and process documents in English and Chinese. Searching in English has not been a difficult problem for HKLII. However, because of the inherent difficulty of recognising word boundaries in Chinese documents, searching in Chinese has been a major challenge for HKLII throughout its development. After trying different methods to improve Chinese searching in HKLII, we have recently got satisfactory results using a new version of the Sino search engine. This paper describes and summarises the experience of HKLII in this regard.

Keywords: HKLII, Chinese Search, bilingual

1. Introduction

1.1. HKLII

Hong Kong Legal Information Institute (HKLII) [1] is a project of the Law and Technology Centre [2], a centre jointly established by the Department of Computer Science [3] and the Faculty of Law of The University of Hong Kong (HKU). As a member of the World Legal Information Institute (WorldLII)[4], HKLII has been greatly assisted by the Australasian Legal Information Institute (AustLII)[5] throughout its development. HKLII is now fully operated by the Department of Computer Science at HKU.

[1] http://www.hklii.hk
[2] http://www.lawtech.hk
[3] http://www.cs.hku.hk
[4] http://www.worldlii.org
[5] http://www.austlii.edu.au

With a view to promoting and supporting the rule of law in Hong Kong, HKLII is a free, independent, non-profit Internet facility providing the public with legal information relating to Hong Kong. Such information includes, among other things, ordinances, regulations, historical laws, judicial decisions, practice directions, domain name arbitration decisions for .hk domain, Privacy Commission's case notes, law reform consultation papers and reports. All information on HKLII is made available to the public free of charge.

The largest set of legal information in the HKLII databases comprises of statutes and judicial decision. Statutes are reproduced from up-to-date data obtained from the Bilingual Laws Information System (BLIS) [6] of the Department of Justice of the Hong Kong Special Administrative Region, and judicial decisions are reproduced from data obtained from the website of the Hong Kong Judiciary[7] on a daily basis. Although the said legal information is already available for free on the websites of BLIS and the Hong Kong Judiciary, HKLII is widely used by the public because of its user-friendly interface and additional features. One of the notable features of HKLII is its ability to conduct universal search across all legal information in Hong Kong.

1.2. BILINGUAL LEGAL INFORMATION

One unique problem that has confronted HKLII since its inception is the bilingual nature of legal information relating to Hong Kong. Up until 1989, statutory laws in Hong Kong were enacted in English only. In April 1989, Hong Kong passed its first bilingual ordinance. Since then, there have been more than 4,000 bilingual ordinances, amending ordinances and pieces of subsidiary legislation enacted in Hong Kong. At present, all statutes in Hong Kong are enacted in both Chinese and English, with both versions enjoying the same legal status.

From 1974 to 1996, the restriction against the use of Chinese in various courts was gradually lifted. In June 1997, the use of Chinese in all civil and criminal proceeding was allowed in the High Court. A bilingual court system was then settled in Hong Kong. While most of the judicial decisions are still written in English, about one fourth of the decisions are now written in Chinese and there is an increasing trend of delivering judicial decisions in Chinese. There is thus a practical need for HKLII to be able to process and search both Chinese and English documents.

[6] http://www.legislation.gov.hk
[7] http://www.judiciary.gov.hk

Previously, HKLII used two search engines, one for Chinese documents and one for English. For English documents, like most other LIIs, HKLII used an earlier version of Sino search engine which was well developed for LIIs to index and search documents in western languages. It was fast, stable and robust. However, this earlier version of Sino did not support indexing and searching non-western languages such as Chinese. Because of this, HKLII had used a General Public Licence search engine called "mnoGoSearch" for indexing and searching Chinese documents. Over the years, we had optimised mnGoSearch for Chinese searching in HKLII using different methods and the result was barely acceptable.

Recently, a new version of Sino was developed with a universal search extension. The purpose of this extension, which requires some pre-processing of documents, is to allow search on any language in UTF-8 encoding using Sino. Since HKLII contains bilingual databases in western language (English) and non-western language (Chinese), HKLII provides a good testing bed for this new version of Sino.

1.3. SINO SEARCH ENGINE

Sino is a high performance free text search engine which aims at speed, flexibility, portability and reliability. Sino, which stands for "Size is no object", exploits the tradeoff between disk space and speed. Generally, the size of the concordance (i.e., the index file) built for a set of documents is about 40% of the total size of the documents. This extra space for indexing results in fast searching by Sino.

Sino consists of two programs, the indexer "Sinomake" and the search engine itself. The normal mode of operation of Sinomake is to rebuild the whole concordance. However, it is possible to invoke Sinomake with extra flags to incrementally update the concordance rather than rebuild it, which is much faster than rebuilding the whole concordance. In HKLII, we build separate concordances for Chinese and English documents, as the indexed words in the two languages are all different.

As for the Sino search engine itself, Sino provides several interfaces for developers to intact with Sino. The most common interface is the Perl Sino API. Sino also has a flexible search parser which supports various logical connectors in search expressions used in different systems such as Google, Lexis and WestLaw.

1.4. mnoGoSearch SEARCH ENGINE[8]

[8] http://www.mnogosearch.org/

mnoGoSearch is a free GPU General Public Licence search engine designed for the Chinese language. It supports Unicode and consists of a built-in dictionary which helps the user to eliminate errors arising from wrong extraction of Chinese words from a document (commonly referred to as "segmentation errors"). mnoGoSearch supports a wide range of databases. In our case, we had chosen to use MySQL, as it was one of the most reliable databases in the open source community.

mnoGoSearch also consists of an indexer and the search engine itself. The indexer of mnoGoSearch basically extracts sentences delimited by punctuation marks, and extracts strings using its built-in dictionary. All extracted strings are stored as indices in MySQL.

In our experience with mnoGoSearch we had encountered two major problems. Firstly, since its dictionary contained only general Chinese terms, many legal terms contained in the Chinese documents of HKLII were not indexed by mnoGoSearch. Secondly, the searching speed of mnoGoSearch was not satisfactory. Searching simple terms might take up to 10 seconds, and searching more complex Boolean queries might take 30 seconds or more. As a result, we had to constantly fine-tune mnoGoSearch in order to provide an acceptable service to our users. This was done until we experimented with the new version of Sino and found it produced satisfactory results.

2. New Sino Search Engine

The new Sino search engine extends the searching capability of its earlier version by introducing a new representation known as "u16a representation" for non-western languages. The new Sino requires converting documents in non-western languages to the new "u16a representation" for indexing and searching. In the case of a Chinese document, this conversion essentially turns all UTF-8 characters in the Chinese document into their alpha-numeric form (ie., their "flat" representation). The objective of this exercise is to enable the Chinese document to be indexed and searched just like any other alpha-numeric document.

2.1. THE U16A REPRESENTATION

The new Sino search engine pre-processes a Chinese document by converting all UTF-8 characters in the document into their hexadecimal form, i.e., "flatten" the character strings into strings consisting only of

digits 0 to 9 and alphabets A to F. However, such a representation may clash with alpha-numeric words in the western language. A good example is the UTF-8 character "香", which in hexadecimal form is "9999", the same as the number "9999" in the western language. To avoid such wrong interpretations, a carefully devised string is appended to each Chinese character in its hexadecimal form to ensure its uniqueness. Based on an analysis of the indexing terms in WorldLII, it is found that the string "u16a" does not appear as an index in the concordance. Using the Google search engine, it is further confirmed that "u16a" is not used in any document written in natural language. Hence the string "u16a" is chosen to be appended to the converted alpha-numeric representation of each Chinese character to create a unique string. This is referred to as the "u16a representation".

As an example, "香港特別行政區" (the Chinese expression for "the Hong Kong Special Administrative Region") in u 16a representation is "9999u16a 6e2fu16a 7279u16a 5225u16a 884cu16a 653fu16a 5340u16a". Each Chinese character is converted into an alpha-numeric string with "u16a" appended to ensure its uniqueness. Once a UTF-8 Chinese document is converted into u16a representation, the new Sino is able to index and search the document in this new representation using the core programs in the earlier version of Sino.

2.2. SHADOW FILE

To preserve the original Chinese documents, the documents in u16a representation are all saved as separate "shadow files". The notion of "shadow files" is a special feature of the Sino search engine. Originally, shadow files were used for indexing non-ascii files (such as pdf files). The idea was to convert non-ascii files into plain text files and let Sino build index on them. These plain text files, with indexed built, were saved as separate files known as "shadow files". Then when a search was requested on the non-ascii files, the shadow files would be searched instead. If the search was successful, the names of the corresponding original non-ascii files would be returned and displayed to the user. Thus by means of such shadow files, Sino was able to search on non-ascii files and return the results to the user.

The new Sino search engine has made use of this notion of shadow files in Chinese searching. As explained above, the u16a representation of a Chinese document is saved separately as a shadow file. Since shadow files are alpha-numeric, Sino is able to index and search their contents. When a user searches for a Chinese key phrase, Sino will convert the key phrase to u16a representation and search the indexes for the

shadow files. If the search is successful, Sino will return the names of the corresponding original files and not the shadow files to the user.

3. Performance analysis of new Sino

In HKLII, there are 14 databases containing, among other things, judgments from various courts, ordinances, regulations, historical laws and other legal information. Since the focus of this paper is searching in Chinese, we have selected, for the purpose of performance analysis of new Sino, 9 databases containing legislation and judicial decisions, which include the largest set of legal information in Chinese.

The machine configuration for conducting the performance analysis was as follows:

Machine: Dell Power Edge R710 Rack Server
Operating System: SunOS 5.10
Processor: Quad-core 2GHz Intel Xeon E5504
RAM: 4GB
Hard disk: 600GB
Apache: 2.0
Perl: 5.8.8

3.1. INDEXING SPEED

The indexing speed of new Sino was measured separately for Chinese and English documents. The result is shown as follows.

	Chinese documents	English documents
Total Number of files	91K	131K
Total File Size	1,272M	1,465M
Time needed for indexing	2m53s	10m51s
Indexing Speed	441M/minute	135M/minute
Size of concordance	396M	862M
Index ratio	31%	59%

The result shows that the indexing speed for Chinese documents is much faster than that for English documents despite that the Total File

Size involved are comparable. This is probably because the number of Chinese characters used in legal documents is relatively limited. The result also shows that the index ratio for Chinese documents is nearly half of that for English documents. This suggests that Chinese characters are repeated more frequently than English words in the documents contained in HKLII.

3.2. UNIQUENESS OF U16A REPRESENTATION

To verify the uniqueness of u16a representation, we randomly chose 20 Chinese names of judges, lawyers, plaintiffs and defendants and searched them in HKLII using new Sino. A total of 491 Chinese documents were returned. We then manually searched for the Chinese names in these Chinese documents. All the Chinese names searched appeared in the Chinese documents returned.

3.3. SEARCH SPEED

To measure the speed of new Sino in Chinese searching, we chose the top 500 Chinese search phrases used by users of our previous search engine mnoGoSearch. These 500 phrases are listed in the Appendix. We also chose to interact with new Sino directly to avoid the interference of network overhead.

3.3.1. Exact Phrase Search

In the exact Phrase Search, we searched for exact occurrences of the 500 chosen phrases in the Chinese documents of HKLII. The average Search time is 0.048 second with an average return of 235 documents.

3.3.2. Freeform Search

In the freeform search (ie., match "any of these words"), we searched for documents which contain any of the Chinese Characters. The average search time is 0.103 seconds with average return of 2,056 documents.

3.3.3. Complex Boolean Expression

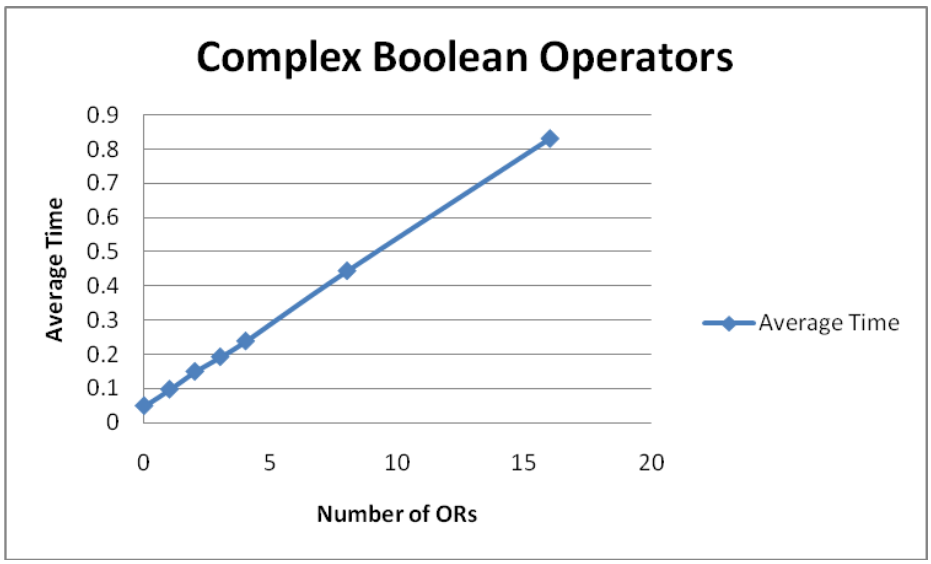
For the complex Boolean expressions, we chose the three common operators "AND", "OR" and "NEAR". We randomly chose two search phrases from the top 500 search phrases and connected them using any of the three operators. Each operator was tested for 1,000 times and the average result was recorded.

Operator	AND	OR	NEAR
Average no. of documents returned	0.997	442	0.560
Average Time	0.096s	0.097s	0.097s

From the above result, it can be seen that all three Boolean operators incurred similar search time. All Boolean expressions used could be searched within very short time.

Since the average number of documents returned for "AND" and "NEAR" was less than one document, we only chose OR to perform more complex queries. We used multiple OR to form more complex search expressions and tested the performance of new Sino. Same as above, we tested for 1,000 times to obtain the average results. The results are shown below.

Number of ORs	Average No of Documents Returned	Average Time
0	235	0.048s
1	442	0.097s
2	708	0.149s
3	907	0.191s
4	1,049	0.238s
8	2,064	0.443s
16	3,692	0.830s



From the above results, it can be seen that for each additional Boolean operator used in the search expression, the search time would increase by about 0.048 seconds. The performance of new Sino was impressive.

4. Future work

4.1. SUPPORT OF SIMPLIFIED CHINESE

The u16a representation described in Section 2 above was tested only on traditional Chinese, as HKLII contains only traditional Chinese documents and English documents. However, as it is foreseeable that users may submit search phrases in simplified Chinese, or that Chinese documents are written or translated to simplified Chinese, the support of simplified Chinese is needed.

4.2. PROBLEM FOR SIMPLIED CHINESE

At present, Chinese documents in HKLII are written in traditional Chinese and accordingly u16a representation shadow files are created for traditional Chinese. But what if a user inputs a search phrase in simplified Chinese? Since all Chinese documents are indexed in traditional Chinese only, a search phrase in simplified Chinese has to be converted to u16a representation for traditional Chinese before conducting the search. However, the mapping of simplified Chinese to traditional Chinese is a "one to many" mapping (eg. the simplified Chinese Character "面" (face) can be converted to "面" (face) or "麵"

(noodles) in traditional Chinese). This means that conversion of search phrases from simplified Chinese to traditional Chinese must be done carefully in order to avoid producing erroneous results.

Conversely, since the mapping of traditional Chinese to simplified Chinese is a "many to one" mapping, two different traditional Chinese Characters may well be mapped to the same simplified Chinese Character. If there are u16a representation shadows files for simplified Chinese, search conducted on these shadow files for search phrases in traditional Chinese may return results which are irrelevant.

In short, the conversion between traditional Chinese and simplified Chinese in both indexing and searching is not as straightforward as it seems and requires further investigation.

5. Conclusion

The new Sino search engine using the u16a representation as outlined in this paper performs well for HKLII. It is fast in both indexing and searching, surpassing the non-western search engines that we had previously encountered. The new Sino search engine has resolved two important problems that HKLII had faced in the past concerning Chinese searching. Firstly, it has avoided the time spent in having to recognise the proper Chinese words contained in the search phrase. As new Sino indexes Chinese documents by character, all search phrases can be handled on the character basis and not on the word basis. Secondly, since u16a representation is alpha-numeric, the new Sino search engine is able to search documents in HKLII in both Chinese and English at the same time.

References

Austin, D. (2000) *Scalability of Web Resources for Law: AustLII's Technical Roadmap: Past, Present and Future*

Pun, K. H. (2003) *Processing Legal Documents in the Chinese-Speaking World: the Experience of HKLII*

Pun, K. H. (2004) *Cross-Referencing for Bilingual Electronic Legal Documents in HKLII*

Mowbray, A. (2008) *Sino – A Text Search Engine*

Appendix

TOP 500 CHINESE SEARCH PHRASES

臨時合約	香港歧視條例
違反合約	离婚
刑罰	警察搜身
個人信貸報告	遺產稅
身份証副本	自僱人仕
自動清盤	有限公司清盤程序
免費律師諮詢服務	法律支援
租印花稅	公眾投訴委員會
香港法治	樓花買賣
欠租收樓	公訴罪行
個人信貸	如何成立業主立案法團
遣散費	公眾地方打架
離婚贍養費	簡易程序
法庭程序	區域法院案件
中國薪俸稅	租務糾紛
案底 求職	信貸資料
標準租約	陪審團
買樓連租約	有限公司清盤
免費法律咨詢	債權人
公司清盤的程序	醫療投訴
遺產管理人	拖欠薪金
租約	分娩假期
計算薪俸稅	家事調解
消費者	臨時買賣合約
法律諮詢免費	聯權共有
香港租務條例	歧視定義
版權	勞工處
平衡進口	侵犯版權條例
刑事訴訟	要約
香港離婚法	公司清盤程序
案底查詢	刑事罪行
物業用途	正式租約
離婚程序	樓宇買賣印花稅
租約期滿	法網
釐印	民事 刑事

大律師公會	合約
商業糾紛	疏忽
免稅額	英國國籍甄選計劃
落口供	破產令解除
歧視	破產令
利得稅	租客不交租
中國個人入息稅	如何成為律師
破產的後果	如何上訴
長命契	個人信貸資料庫
離婚 財產分配	拘捕令
離婚問題	應評稅利潤
職業安全及健康條例	香港 普通法
租約厘印費	刑事案件
離婚後財產	少年罪犯
無律師代表訴訟人資源中心	香港大律師公會
呈請書	破產管理署
法定病假	民事訴訟期限
分租	身分證副本
香港法網	遺產執行人
應課薪俸稅	家事法庭離婚
香港法院	民事訴訟
共同撫養權	臨時買賣合約樣本
投訴 醫生	當值律師
香港律師收費	保險
租約打厘印	清盤令
英國公民護照	勞資審裁處案例
永久性居民	銘謝
法定要求償債書	厘印費計算
可公訴罪行	性別歧視
自僱人士合約樣本	如何取回強制性公積金
釐印費計算	香港過期居留
勞工處法定假期	聆案官
證人陳述書	香港國際仲裁中心
外國註冊結婚	不小心駕駛 刑事
自僱合約	香港離婚財產分配
香港永久居民身份資格	民事法
高等法院案件	共同申請書
刑事恐嚇	香港工作簽證
香港土地買賣	分權共有

遺產承辦手續	庭外和解
不合法解僱	終止僱傭合約
勞資糾紛	社區法律
象徵式	引導性問題
家庭崗位歧視	債務重組
社區法網	私隱專員公署
消費者委員會 投訴熱線	香港律師公會
離婚訴訟程序	民事訴訟案例
租約問題	租務
樓宇臨時買賣合約	免費法律
私隱定義	查案底
有薪病假	香港醫務委員會
信貸紀錄	物業印花稅
免費法律諮詢電話	投訴診所
訂立遺囑	厘印費
刑事責任	賣買合約
民事刑事	社區法
保險箱	離婚申請書
信貸資料服務機構	離婚財產分配
破產 後果	產假
因工受傷	續租
贍養費	謹慎責任
香港土地法	利得稅計算
物業買賣程序	普通法 香港
正式買賣合約	租客
家庭崗位歧視條例案例	小額錢
自雇人士	律師費用的計算
個人資料	法定假期
疾病津貼	版權有效期
身分證號碼	離婚影響
遺囑認證書	租客欠租
醫務委員會投訴	沒有遺囑
香港普通法	貨品售賣條例
小 額 錢 債 審 裁 處	物業稅
香港刑事法	離婚財產如何分配
離婚申請	連續性合約
香港醫務委員會 投訴	利得稅免稅額
抗辯書樣本	遺產分配
賣樓印花稅	法律援助

物業厘印費	投訴牙醫
公訴程序	如何立遺囑
身份證號碼	個人信貸資料
陪審員資格	年假
物業臨時買賣合約	香港法律制度
厘印	清盤程序
誓詞	如何計算印花稅
勞資糾紛處理	如何申請離婚
離婚財產	香港勞工處
樓宇印花稅	薪俸稅 計算
雙重國籍	外籍家庭傭工
酒牌局網址	病假
法庭案件	香港免費法律諮詢
利得稅稅率	樓宇臨時買賣合約樣本
租約厘印費計算	香港法律來源
抗辯書	個人自願安排
樓宇買賣程序	租務條例
仲裁	擾人清夢
離婚	遺產承辦處
消費者委員會投訴過程	香港遺產法
警察權力	僱傭合約
父母免稅額	大廈公契
業主收樓	投訴醫生
平機會 懷孕 實例	個人信貸紀錄
香港法律	民事訴訟程序
樓宇厘印費計算表	律師
申請將暫准判令轉為絕對判令通知書	摸貨
衡平法	律師公會
業主立案法團	信貸資料機構
樓宇買賣	毀謗
如何訴訟	訟費評定
刑事紀錄	遺產承辦
成為律師	免費法律意見
供養父母免稅額	釐印費
臨時賣買合約	公契
雙糧	律政司
個人免稅額是多少	香港律師公會網頁
個人資料定義	事務律師
無遺囑遺產分配	中國入息稅

強制性公積金	破產
保險公司	遺產承辦處地址
租	民事
自僱人士合約	醫療疏忽
家事法庭離婚審訊	反歧視條例
司法覆核	可保權益
銀行保險箱	家庭暴力
案底	業主與租客
香港法庭案件	連續性僱傭合約
殘疾歧視	分居協議
頂手費	保險冷靜期
醫務委員會	家事法庭登記處表格 5a
物業買賣	工廠及工業經營條例
傳訊令狀	遺產
年終酬金	公司清盤
撫養權	小額錢債
樓宇買賣合約	免費法律諮詢計劃
印花稅計算方法	民事訴訟費用如何算
自僱人士	律師費用
印花稅計算	醫務委員會 投訴
醫療事故	租樓 注意
薪俸稅免稅額	律師費
病假工資	社區
普通法 衡平法	長期服務金
樓宇買賣律師費	遺產承辦處
個人入息課稅	遺 產 承 辦 處
殘疾歧視條例	年終酬金計算
買賣樓宇	香港的法律制度
買賣合約	家事法庭登記處
香港永久性居民	傷殘受養人免稅額
文事訴訟	已婚人士免稅額
刑事法	離婚 財產
陪審員	財產分配
免費法律諮詢	買賣交易
租樓注意事項	租樓注意
自僱合約樣本	個人資料 定義
保險 冷靜期	法律援助輔助計劃
厘印費用	租務問題
僱傭合約樣本	香港 雙重國籍

刑事檢控程序	遺囑
商業合約	香港遺產承辦處
誹謗	法律諮詢
香港雙重國籍	獨資經營
工作簽證	工傷賠償
身份證副本	租約釐印費
免費法律諮詢服務	如何解除破產令
自動清盤程序	產假薪金
勞工法例	殘疾歧視定義
侵犯版權	遺產管理書
小額錢債審裁處	不合理解僱
申請遺產承辦書	樓花
誓章	印花稅
有薪假期	勞資
臨時買賣合約範本	長期服務金 辭職
租務法例	有用網站
抗辯	醫管局投訴
應課薪俸稅入息	個人私隱定義
樓宇買賣 律師費	薪俸稅
收樓	勞資審裁處
物業買賣利得稅	臨時租約
婚姻問題	子女贍養費
地產代理佣金	解除破產令
買賣樓宇程序	罪犯自新條例
標準稅率	香港法律的來源
連租約	反申索書
離婚呈請書	大律師
消費者投訴	香港租務法例
租賃印花稅	醫療失誤
報稅計算	子女撫養權
大律師 事務律師	反歧視
法律意見	破產後果
續租權	刑事民事
遺產處理	當值律師計劃
長期服務金計算方法	申請離婚
誹謗條例	新租務條例
樓宇買賣合約樣本	版權條例
刑事 民事	連續性的僱傭合約
離婚後財產分配	遺囑執行人

刑罰種類	醫院管理局投訴
住宅用途	免費法律服務
遺產承辦書	簡易程序罪行
平機會地址	欠租
如何提出民事訴訟	小額錢債審裁處條例
小額錢債審裁處查詢	樓宇買賣費用
離職聲明	銀行擔保
民事索償	家事法庭
長期服務金計算	印費
律師 收費	連租約買賣
家庭崗位	冷靜期
普通法	如何計算薪俸稅
離婚的影響	大訂
立遺囑	當值律師服務
小額錢債審裁處地址	解除破產
投訴醫院	離婚證書表格
追討欠租	遺產繼承
律師收費	厘印費計算表
初級偵訊	清盤
刑事案底	勞工糾紛