Considerations for an interactive analytical search procedure in a large text corpus of premodern Chinese texts

Christian Wittern

Introduction

Over the last decades, a tremendous amount of premodern Chinese has become available in openly accessible online databases, e.g. *The Chinese Text Project* (See https://ctext.org), *Kanseki Repository* (https://kanripo.org) and *Wikisource* (https://zh.wikisource.org).

This wealth of data poses the challenge of exploring this textual universe in a meaningful way. A researcher should be able to discover when terms appeared, in what context the meaning developed and what changes can be observed. In this paper, we describe the implementation of a search function for such a corpus of documents, which maps the results across a hierarchical taxonomy of features of different dimensions. One dimension for example would be the creation date of the document, placed in a hierarchy of date brackets, that are meaningful to researchers of the history of Chinese language. Another dimension would be the classified catalog used in traditional Chinese bibliography; this allows the user to identify hotspots and clusters of terms and compare them across bibliographical categories. The dimensions can also be applied sequentially, in order to narrow done the scope of investigation. Search results can then be used to create new sets of texts, which can themselves placed in a hierarchical taxonomy defined by the user, thus creating a powerful tool for iterative knowledge discovery.

The search engine is part of a larger project for collaborative research on premodern Chinese materials, which provides tools for close reading, text critical editing, linguistic annotation and translation. The annotations created through these activities do provide additional dimensions for the search engine and allow the researchers to further customize the result according to their needs through selecting and narrowing or widening the scope.

This presentation will look at ways the search results can be evaluated against a classified hierarchical taxonomy.

What is the TLS / 漢學文典 Hànxué Wéndiǎn?

The Thesaurus Linguae Sericae (TLS) is a comprehensive database for the study of premodern Chinese, begun by Christoph Harbsmeier at Oslo University in the 1980s. It was implemented by Jens Østergaard Petersen (Copenhagen) over a long period of time using the Filemaker database application.

TLS is a very ambitious project, not only trying to systematically analyse the development and structure of Chinese conceptual schemes, but to do this in a comparative perspective. The meta-theoretical framework for this is very much inspired by the approach to systematic lexicography put forward by Juri Apresjan (see Apresjan 2000 for a good introduction). The implementation details and focus changed a lot over the years, following the interests of the principal investigator, but the

project has always been conceived as a collaborative investigation into the Chinese language and thought and has been able to ensure participation of many high profile scholars in China and the West.

This tension between the theoretical aspirations and the actual technical implementation in a single-user commercial database application has characterized the project for a long time. There have been many attempts to overcome this, and since 2009 the TLS has been available on the Internet, although only has a static snapshot of the database without updating functions.

The database is now developed at the **Center for Informatics in East-Asian Studies**, Institute for Research in Humanities, Kyoto University, with support from the Dean for Research, Department of East Asian Studies, and Program in East Asian Studies, Princeton University. Development is also supported by the **sin-aps research group** at Friedrich-Alexander-Universität Erlangen Nürnberg, with financial support from the Alexander von Humboldt Foundation.

It is hosted in cooperation with Ruhr University Bochum, **Center for the Study of Traditional Chinese Cultures** . Support from Heidelberg University - Cluster of Excellence - Asia and Europe in a Global Context and IKOS - University of Oslo is also gratefully acknowledged.

Overview of the TLS

A more thorough introduction has been given in Wittern (2021), so this will be just a short summary. In some ways, the TLS can be seen as a textual database with extensive features for annotation of texts. In the first decades of the project, the establishing of a philological trustworthy text has been outside the scope of the TLS. So texts that had already been established and punctuated form the core of the TLS, with a focus on the main philosophical and historical texts up to the Han. Buddhist texts translated into Chinese, as well as indigenous Buddhist texts, however, have also been part of the database for many years. The texts are displayed in short lines that form somehow as smallest semantic units, usually less than 10 characters of length.

The are different types of annotations, but the most important and frequent ones are lexical annotations, that identify the syntactic and/or semantic function of a word (lexeme) and associate this with a "Concept", which serves as an intermediate ordering layer for the core semantic field of a word.

Concepts can thus be understood as grouping together words that can act as synonyms or near synonyms for a given semantic field. They are themselves also defined using a logical description that form a coherent conceptual space in a Leibnizian sense.

The annotations themselves thus form the bridge between the abstract conceptual space, with a taxonomy of concepts, and the texts, where the concepts are observed.

In addition to purely linguistic annotations, a considerable effort has also made to describe and mark rhetorical devices (using the terminology developed in European rhetoric) as they are employed by Chinese authors.

漢學文典 Hànxué Wéndiǎn

漢學文典 Hànxué Wéndiǎn (HXWD) is the Chinese name of the TLS database and also the title of the new collaborative website, which is available at hxwd.org. Below, I will use HXWD when referring specifically to this new website, but continue to use TLS to refer to the whole project in a more general sense.

The new website has been designed to support collaborative research from the outset. The main

features and technical characteristics are as follows:

- The texts are encoded using XML files with markup as recommended by the `Text Encoding Inititiative` (TEI, see tei-c.org)

- Annotations and other descriptive notes are represented using `XML`, based on TEI markup, but with modifications for specialized constructs.

- The `eXist XML database` is used in the backend, with `XQuery` to serve the content to the browser.

- The user interface in the browser is based on `bootstrap` and `jQuery` and `Javascript`.

  Registered users can add translations, annotations and (depending on the individual permissions) also add or modify concepts, phonetic profiles, character taxonomy and many other types of information. The website is still under development, but we started to accept requests for user accounts recently.

The question: classified hierarchy and interactive search

Classified bibliographical catalogs have for a long time been a feature of Chinese bibliography (See for example Inami Ryoichi 2003, p11-68), by the Tang the division in four main sections and most major sections have been established, while the subsections to have a certain amount of variation until the Siku Quanshu in 1778. The Kanseki Repository, on which the taxonomy used here is based, has further developed and harmonized this (see Wittern 2016, p22-29). For the current project, the structure of the catalog has been again amended, as a third level has been introduced for categories with many texts for pragmatic reasons, to allow more specific queries being processed faster. Appendix A shows the current catalog, together with the number of texts in a category, among other things.

The catalog maps the texts into a hierarchical taxonomy of semantically distinct domains. Used in interactive search to intersect the set of matching phrases with this taxonomy enables the researcher to gauge the relevance of the search term relative to these different sections. Since the number of characters in these sections differs widely, the absolute number of hits does little to show the provenance of search terms to specific sections.

This report is a first attempt at trying some avenues to get a better understanding of possible ways to meaningfully discover the productive discrepancy between the number of hits that would be expected for a certain section due to the size of that section and the observed number of hits. There are a number of different possible approaches to this questions and below, I am trying a few. One of the difficult problems about working on this is the question of how to evaluate the results. There seems to be no measurement to evaluate the results except for inspecting and comparing a sufficient number of samples.

TF/IDF, TF/ICF and beyond: arriving at an appropriate algorithm

The most common algorithm to score relevance against text size seems to be the term frequency / inverse document frequency, which simply offsets the frequency of terms by the number of documents containing the term against the total number of documents in the corpus.

$$\mathbf{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Figure 1: TF

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Figure 2: IDF

Figure 1 gives a frequently used way to calculate term frequency, by dividing the observed frequency f of a term t in document d by the sum of these values over the whole set of documents.

The inverse document frequency here is calculated as the logarithmically scaled inverse fraction of the number of documents containing a term. In Figure 2 N is the total number of documents, which is divided by the number of documents containing the term. Sometimes 1 is added to this to ensure that division by zero is avoided, in the present context this is not necessary, since we are only looking at matches that are actually in the set of documents.

In the current context, both term frequency and inverse document frequency needs to be reconsidered. The issue about TF is that we can not naïvely use terms here. While the terms searched for will usually be terms in the narrow sense of the term, we can not count terms without tokenizing the text. The easiest measure of token counts would be characters, but in a way counting characters and comparing this with matches of multi-term characters seems like comparing apples and oranges.

Inverse document frequency also has its problems. Since the aim here is not discovery of documents in a corpus of documents with similar length, the measurement seems not to be well suited to this task. On top of that in the context of the database where this measurement is to be used for interactive search, it turned out that getting the number of documents matched is prohibitively expensive, so it has to be adjusted for the task at hand.

A different approach is proposed by Reed et.al (2006), which uses ICF (inverse corpus frequency) instead of IDF. In this approach, a static table of document count per term is used instead of calculating a value for corpora of different sizes, the values than could simply looked up in a table rather than calculated on demand. To realize this approach with our corpus, there is still the question of what tokens to use, which again would have to be ngrams of the appropriate size.

$$w_{ij} = \log\left(1 + f_{ij}\right) \times \log\left(\frac{N+1}{n_j + 1}\right)$$

Figure 3: TF/ICF

To overcome these problems and adopt them to the text corpus at hand, a modified algorithm has been developed as follows:

- term frequency is calculated as count of matches in a given section against the number of

characters in that section. Although that does not provide a satisfying solution, it is easy to calculate and provides a good-enough measure for the problem at hand. In order to clearly communicate the fact that this is not the term frequency, this will be called match frequency, MF.

- The significance of the matches is calculated against the number of phrases containing the match against the total number of phrases. Conceptually, this can be seen as defining a document as consisting of exactly one phrase. This also takes into account the fact that the phrase is the unit of interest when looking for matches and it is easily calculated.

Since the phrase is called internally **seg(ment)** in the database, this algorithm will be called MF/ISF, match frequency vs inverse segment frequency. For the sake of comparison, the inverted corpus frequency based on the total number of characters in each section is also given in the tables below, this is marked as ICF for inverse character frequency and MF/ICF for the combined value.

Experiments with the algorithm

In practice, the value found for all of the subsections will be compared to determine the relative uniqueness of the matches in a given section. The deviation from the expected value will indicate, how specific a term is in a given section. In the search interface, for the purpose of clearly communicating the outliers to the user, a cut-off has been set to render the value in different colors, as shown in Figures 4, 5 and 6. However, again, a proper value of this cut-off is not easily to gauge in advance, so the user is encouraged to experiment and set an appropriate cut-off value dependant on her needs. The initial value set by the system is at +/- 0.2, which seems appropriate for many cases. The interface also offers a matrix with all values, to give the user the possibility to evaluate the results differently. For the results displayed in the interface, only the categories with matching results are shown, the others are not displayed.

For the experiments reported here, a set of 7587 texts has been used. The distribution of these texts across the hierarchy is shown in Appendix A, which gives text counts, number of characters and number of segments for all categories and sub-categories. For the purpose of these experiments, the four traditional categories 四部 (經部, 史部, 子部 and 集部) have been grouped together, to be contrasted with the additional Daoist 道部 and Buddhist 佛部 texts. The levels of comparison can be seen in the screenshots from the system (which have been arranged in two columns to fit on the page) by the indent: categories on the same level are aligned with the same amount of indent.

As the system is still in an experimental state and other calculations will be explored, for each search conducted a matrix table with the MF/ICF and MF/ISF values are stored. For the examples discussed here, the matrix tables are given in Appendix B.
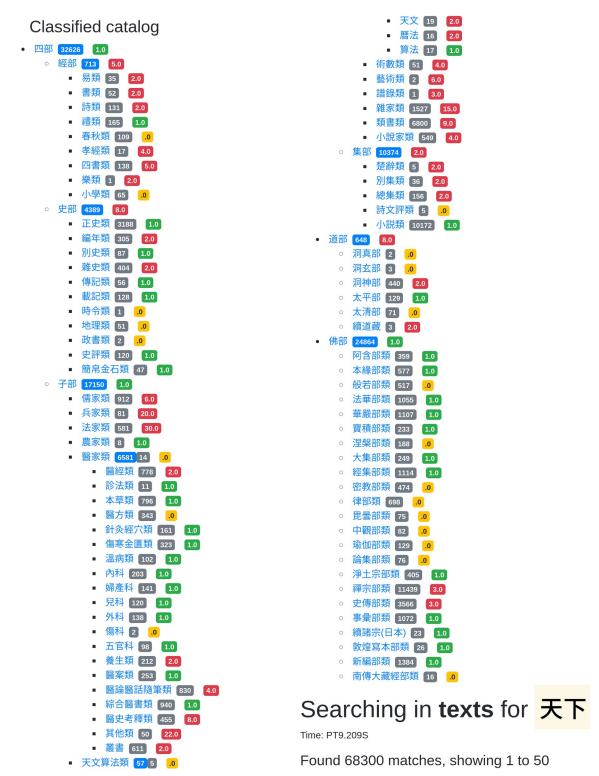
Classified catalog

- 四部 `32626` `1.0`
  - 經部 `713` `5.0`
    - 易類 `35` `2.0`
    - 書類 `52` `2.0`
    - 詩類 `131` `2.0`
    - 禮類 `165` `1.0`
    - 春秋類 `109` `.0`
    - 孝經類 `17` `4.0`
    - 四書類 `138` `5.0`
    - 樂類 `1` `2.0`
    - 小學類 `65` `.0`
  - 史部 `4389` `8.0`
    - 正史類 `3188` `1.0`
    - 編年類 `305` `2.0`
    - 別史類 `87` `1.0`
    - 雜史類 `404` `2.0`
    - 傳記類 `56` `1.0`
    - 載記類 `128` `1.0`
    - 時令類 `1` `.0`
    - 地理類 `51` `.0`
    - 政書類 `2` `.0`
    - 史評類 `120` `1.0`
    - 簡帛金石類 `47` `1.0`
  - 子部 `17150` `1.0`
    - 儒家類 `912` `6.0`
    - 兵家類 `81` `20.0`
    - 法家類 `581` `30.0`
    - 農家類 `8` `1.0`
    - 醫家類 `6581` `14` `.0`
      - 醫經類 `778` `2.0`
      - 診法類 `11` `1.0`
      - 本草類 `796` `1.0`
      - 醫方類 `343` `.0`
      - 針灸經穴類 `161` `1.0`
      - 傷寒金匱類 `323` `1.0`
      - 溫病類 `102` `1.0`
      - 內科 `203` `1.0`
      - 婦產科 `141` `1.0`
      - 兒科 `120` `1.0`
      - 外科 `138` `1.0`
      - 傷科 `2` `.0`
      - 五官科 `98` `1.0`
      - 養生類 `212` `2.0`
      - 醫案類 `253` `1.0`
      - 醫論醫話隨筆類 `830` `4.0`
      - 綜合醫書類 `940` `1.0`
      - 醫史考釋類 `455` `8.0`
      - 其他類 `50` `22.0`
      - 叢書 `611` `2.0`
    - 天文算法類 `57` `5` `.0`
      - 天文 `19` `2.0`
      - 曆法 `16` `2.0`
      - 算法 `17` `1.0`
    - 術數類 `51` `4.0`
    - 藝術類 `2` `6.0`
    - 譜錄類 `1` `3.0`
    - 雜家類 `1527` `15.0`
    - 類書類 `6800` `9.0`
    - 小說家類 `549` `4.0`
  - 集部 `10374` `2.0`
    - 楚辭類 `5` `2.0`
    - 別集類 `36` `2.0`
    - 總集類 `156` `2.0`
    - 詩文評類 `5` `.0`
    - 小説類 `10172` `1.0`
- 道部 `648` `8.0`
  - 洞真部 `2` `.0`
  - 洞玄部 `3` `.0`
  - 洞神部 `440` `2.0`
  - 太平部 `129` `1.0`
  - 太清部 `71` `.0`
  - 續道藏 `3` `2.0`
- 佛部 `24864` `1.0`
  - 阿含部類 `359` `1.0`
  - 本緣部類 `577` `1.0`
  - 般若部類 `517` `.0`
  - 法華部類 `1055` `1.0`
  - 華嚴部類 `1107` `1.0`
  - 寶積部類 `233` `1.0`
  - 涅槃部類 `188` `.0`
  - 大集部類 `249` `1.0`
  - 經集部類 `1114` `1.0`
  - 密教部類 `474` `.0`
  - 律部類 `698` `.0`
  - 毘曇部類 `75` `.0`
  - 中觀部類 `82` `.0`
  - 瑜伽部類 `129` `.0`
  - 論集部類 `76` `.0`
  - 淨土宗部類 `405` `1.0`
  - 禪宗部類 `11439` `3.0`
  - 史傳部類 `3566` `3.0`
  - 事彙部類 `1072` `1.0`
  - 續諸宗(日本) `23` `1.0`
  - 敦煌寫本部類 `26` `1.0`
  - 新編部類 `1384` `1.0`
  - 南傳大藏經部類 `16` `.0`

Searching in **texts** for 天下

Time: PT9.209S

Found 68300 matches, showing 1 to 50

Figure 4: Search for 天下

The results for a fair number of searches have been recorded, but here it will suffice to show three representative samples. The first one is for the term 天下, 'all under Heaven', a term that has political as well as philosophical connotations, so one would expect high values in the categories dealing with philosophy or history, but not so much for Buddhism, medical or mathematical texts. The display that would be shown to users with the default settings is shown in Figure 4, as can be seen it shows few matches for medical texts, but more than expected for the Buddhist texts and also

an unexpected high turnout in the literary section (集部).

A completely different case is 半夏, the name for the crow dipper (pinellia ternata), the root of which is a frequently used medical agent in traditional Chinese medicine. But it also has a literal meaning, 'half of summer', so result matches will not be limited to the medical use. Figure 5 shows the results of this search, again there are some unexpected peaks, in the Buddhist categories dealing with Vinaya, Buddhist history and Chan. This will have to be further investigated, but maybe suggests a usage related to periods of intensive retreats in summer.
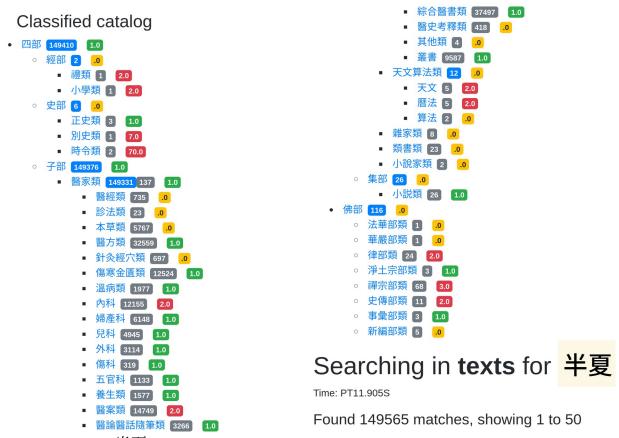
Classified catalog

- 四部 149410 1.0
  - 經部 2 .0
    - 禮類 1 2.0
    - 小學類 1 2.0
  - 史部 6 .0
    - 正史類 3 1.0
    - 別史類 1 7.0
    - 時令類 2 70.0
  - 子部 149376 1.0
    - 醫家類 149331 137 1.0
      - 醫經類 735 .0
      - 診法類 23 .0
      - 本草類 5767 .0
      - 醫方類 32559 1.0
      - 針灸經穴類 697 .0
      - 傷寒金匱類 12524 1.0
      - 溫病類 1977 1.0
      - 內科 12155 2.0
      - 婦產科 6148 1.0
      - 兒科 4945 1.0
      - 外科 3114 1.0
      - 傷科 319 1.0
      - 五官科 1133 1.0
      - 養生類 1577 1.0
      - 醫案類 14749 2.0
      - 醫論醫話隨筆類 3266 1.0
      - 綜合醫書類 37497 1.0
      - 醫史考釋類 418 .0
      - 其他類 4 .0
      - 叢書 9587 1.0
    - 天文算法類 12 .0
      - 天文 5 2.0
      - 曆法 5 2.0
      - 算法 2 .0
    - 雜家類 8 .0
    - 類書類 23 .0
    - 小說家類 2 .0
  - 集部 26 .0
    - 小說類 26 1.0
- 佛部 116 .0
  - 法華部類 1 .0
  - 華嚴部類 1 .0
  - 律部類 24 2.0
  - 淨土宗部類 3 1.0
  - 禪宗部類 68 3.0
  - 史傳部類 11 2.0
  - 事彙部類 3 1.0
  - 新編部類 5 .0

Searching in **texts** for 半夏

Time: PT11.905S

Found 149565 matches, showing 1 to 50

Figure 5: Search for 半夏

The third term used for demonstration here is 國語, the 'National Language', but also the title of a collection of anecdotes compiled in the late Warring States period. As it turned out, many of the matches shown in Figure 6 show locations where a quote from this text is used. While not a question of the appropriateness of the algorithm, to make this functionality useful for its purpose, it has to be considered how to deal with such cases.
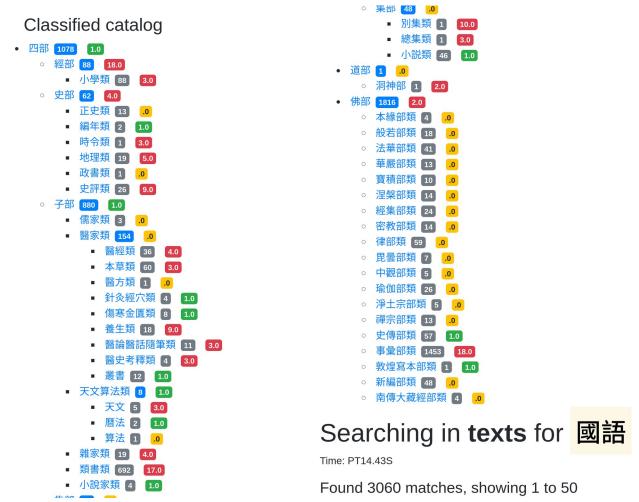
Classified catalog

- 四部 `1078` `1.0`
  - 經部 `88` `18.0`
    - 小學類 `88` `3.0`
  - 史部 `62` `4.0`
    - 正史類 `13` `.0`
    - 編年類 `2` `1.0`
    - 時令類 `1` `3.0`
    - 地理類 `19` `5.0`
    - 政書類 `1` `.0`
    - 史評類 `26` `9.0`
  - 子部 `880` `1.0`
    - 儒家類 `3` `.0`
    - 醫家類 `154` `.0`
      - 醫經類 `36` `4.0`
      - 本草類 `60` `3.0`
      - 醫方類 `1` `.0`
      - 針灸經穴類 `4` `1.0`
      - 傷寒金匱類 `8` `1.0`
      - 養生類 `18` `9.0`
      - 醫論醫話隨筆類 `11` `3.0`
      - 醫史考釋類 `4` `3.0`
      - 叢書 `12` `1.0`
    - 天文算法類 `8` `1.0`
      - 天文 `5` `3.0`
      - 曆法 `2` `1.0`
      - 算法 `1` `.0`
    - 雜家類 `19` `4.0`
    - 類書類 `692` `17.0`
    - 小說家類 `4` `1.0`
    - 集部

- 集部 `48` `.0`
  - 別集類 `1` `10.0`
  - 總集類 `1` `3.0`
  - 小説類 `46` `1.0`
- 道部 `1` `.0`
  - 洞神部 `1` `2.0`
- 佛部 `1816` `2.0`
  - 本緣部類 `4` `.0`
  - 般若部類 `18` `.0`
  - 法華部類 `41` `.0`
  - 華嚴部類 `13` `.0`
  - 寶積部類 `10` `.0`
  - 涅槃部類 `14` `.0`
  - 經集部類 `24` `.0`
  - 密教部類 `14` `.0`
  - 律部類 `59` `.0`
  - 毘曇部類 `7` `.0`
  - 中觀部類 `5` `.0`
  - 瑜伽部類 `26` `.0`
  - 淨土宗部類 `5` `.0`
  - 禪宗部類 `13` `.0`
  - 史傳部類 `57` `1.0`
  - 事彙部類 `1453` `18.0`
  - 敦煌寫本部類 `1` `1.0`
  - 新編部類 `48` `.0`
  - 南傳大藏經部類 `4` `.0`

Searching in **texts** for 國語

Time: PT14.43S

Found 3060 matches, showing 1 to 50

Figure 6: Search for 國語

Conclusions

An adoption of the widely used TF/IDF has been introduced, with slight changes to how term frequency and document occurrence are calculated. This has been used dynamically to evaluate search results against a classified hierarchy of categories, under which the texts are organized. The purpose is mainly discovery of unexpected occurrences or meanings of the given search term. To aid in the quick evaluation, the results have been divided in three categories of "higher number of matches than expected" (marked in red), "in the expected range" (marked green) and "lower than expected" (marked in yellow), with an arbitrarily chosen cut-off value.

The question of whether this approach is actually useful to the users is not yet clear, since the details of the algorithm and the cut-off value have not yet been finalized. Initial tests with some users provided positive feedback and the explicit wish to have this as a permanent feature of the system. For further evaluation in the future a feedback mechanism might be provided, it might also be possible to provide the users with a way to e.g. set their own cut-off value.

Overall, the approach of using a hierarchical taxonomy to analyse search results looks promising and might in the future also be extended to other types of taxonomy, for example the taxonomy of dates for the historical development of the Chinese language, which is already in use in the database.

References

Apresjan, Juri. Systematic Lexicography. 2000.

Inami Ryoichi (井波陵一). 知の座標, Kyoo 2003.

Reed, Joel, Yu Jiao, Thomas Potok, Brian Klump, Mark Elmore, and Ali Hurson. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams. Proceedings - 5th International Conference on Machine Learning and Applications, ICMLA 2006, 2006.

Wittern, Christian (ed.) CIEAS Research Report 2015: The Kanseki Repository, Kyoto 2016

Wittern, Christian. KanripoX: A tagset for connecting digital texts, ORICOM 2021, http://kanji.zinbun.kyoto-u.ac.jp/seminars/oricom/PDFs/2021PDFs/2wittern.pdf .

Appendix A

Categories of the classified taxonomy and the values for each category. Categories with gray background are top level categories, the other subcategories; the level can be read from the length of the category ID value.

ID	Name	# of texts	# of characters	# of segments
	Total	7587	580159200	93428355
notav	unassigned	969	37241125	4698254
KR1	經部	26	2097906	284478
KR1a	易類	1	21792	7413
KR1b	書類	2	42072	9539
KR1c	詩類	2	92353	21768
KR1d	禮類	4	243044	57480
KR1e	春秋類	6	369052	90059
KR1f	孝經類	2	6608	1613
KR1h	四書類	2	51539	11539
KR1i	樂類	1	985	217
KR1j	小學類	6	1270461	84850
KR2	史部	93	5190208	992120
KR2a	正史類	6	3553021	680996
KR2b	編年類	2	192637	38723
KR2d	別史類	2	105426	22077
KR2e	雜史類	2	198304	39727
KR2g	傳記類	3	85324	17419
KR2i	載記類	5	210912	42480
KR2j	時令類	1	21955	4736
KR2k	地理類	3	340280	60200
KR2m	政書類	1	182708	33956
KR2o	史評類	2	267446	44316
KR2p	簡帛金石類	66	32195	7490
KR3	子部	1468	281112945	51987145
KR3a	儒家類	15	2155062	450992
KR3b	兵家類	10	48961	12141
KR3c	法家類	4	270798	59680
KR3d	農家類	1	82774	20367
KR3e	醫家類	9	398385	93941
KR3ea	醫經類	55	15082039	2889003
KR3eb	診法類	18	761404	149049
KR3ec	本草類	76	29031883	5540436
KR3ed	醫方類	146	41278769	7943869
KR3ee	針灸經穴類	54	8345539	1584618
KR3ef	傷寒金匱類	85	19871955	3713949
KR3eg	溫病類	48	4296369	749887
KR3eh	內科	56	13896149	2508655
KR3ei	婦產科	73	9165217	1701554
KR3ej	兒科	58	8788518	1629630
KR3ek	外科	50	9887099	1805219
KR3el	傷科	18	484885	86081

KR3em	五官科	39	3455919	597846
KR3eo	養生類	45	3492131	644275
KR3ep	醫案類	103	12766211	2297300
KR3eq	醫論醫話隨筆類	97	7743177	1357983
KR3er	綜合醫書類	112	50051836	9201557
KR3es	醫史考釋類	6	2164217	397417
KR3et	其他類	2	75236	16243
KR3eu	叢書	61	16223211	2888303
KR3f	天文算法類	16	311268	29533
KR3fa	天文	38	1578251	100422
KR3fb	曆法	24	1089958	85572
KR3fc	算法	79	2678699	252704
KR3g	術數類	4	134174	34706
KR3h	藝術類	4	4322	1087
KR3i	譜錄類	1	3231	899
KR3j	雜家類	24	1476632	304080
KR3k	類書類	22	12020887	2415037
KR3l	小說家類	15	1997779	423110
KR4	集部	340	60219600	8860721
KR4a	楚辭類	1	15271	2596
KR4b	別集類	4	104996	19035
KR4h	總集類	3	315238	64846
KR4i	詩文評類	3	44549	9394
KR4k	小説類	329	59739546	8764850
KR5	道部	19	989026	121551
KR5a	洞真部	2	13009	2758
KR5b	洞玄部	3	9430	1791
KR5c	洞神部	11	592493	48915
KR5e	太平部	1	218256	40998
KR5f	太清部	1	154434	26727
KR5h	續道藏	1	1404	362
KR6	佛部	4672	193308390	26484086
KR6a	阿含部類	181	2780203	538990
KR6b	本緣部類	87	2669831	472621
KR6c	般若部類	229	7984695	1198086
KR6d	法華部類	239	14811978	1776182
KR6e	華嚴部類	166	12512314	1343983
KR6f	寶積部類	110	2746790	431476
KR6g	涅槃部類	54	3692972	484468
KR6h	大集部類	38	1674153	263389
KR6i	經集部類	630	13369291	1775806
KR6j	密教部類	731	13279605	1471364
KR6k	律部類	266	15440115	2403106
KR6l	毘曇部類	40	6892837	985161
KR6m	中觀部類	44	1714247	261196
KR6n	瑜伽部類	135	9741442	1312180
KR6o	論集部類	120	3220292	444403
KR6p	淨土宗部類	141	3296147	465023
KR6q	禪宗部類	610	33993470	4637630
KR6r	史傳部類	197	9722423	1499204

KR6s	事彙部類	145	9497025	1185741					
KR6t	續諸宗(日本)	2	443966	44964					
KR6u	敦煌寫本部類	57	192030	26370					
KR6v	新編部類	367	16637195	2309123					
KR6w	南傳大藏經部類	83	6995369	1153620					

Appendix B

Matrix tables for the three search examples.

1. Searching for 天下

Total number of hits: 68300

Category		Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR0	四部	0	32626		5.61E-01	6.42E-01	8.74E-01	7.00E-01	8.02E-01
KR1	經部	0	713		2.19E-02	6.02E-03	3.63E+00	4.58E-03	4.77E+00
KR1a	易類	1		35	4.91E-02	1.04E-02	4.73E+00	2.61E-02	1.88E+00
KR1b	書類	2		52	7.29E-02	2.01E-02	3.64E+00	3.35E-02	2.18E+00
KR1c	詩類	2		131	1.84E-01	4.40E-02	4.17E+00	7.65E-02	2.40E+00
KR1d	禮類	4		165	2.31E-01	1.16E-01	2.00E+00	2.02E-01	1.15E+00
KR1e	春秋類	6		109	1.53E-01	1.76E-01	8.69E-01	3.17E-01	4.83E-01
KR1f	孝經類	2		17	2.38E-02	3.15E-03	7.57E+00	5.67E-03	4.21E+00
KR1h	四書類	2		138	1.94E-01	2.46E-02	7.88E+00	4.06E-02	4.77E+00
KR1i	樂類	1		1	1.40E-03	4.70E-04	2.99E+00	7.63E-04	1.84E+00
KR1j	小學類	6		65	9.12E-02	6.06E-01	1.51E-01	2.98E-01	3.06E-01
KR2	史部	0	4389		1.35E-01	1.49E-02	9.04E+00	1.60E-02	8.42E+00
KR2a	正史類	6		3188	7.26E-01	6.85E-01	1.06E+00	6.86E-01	1.06E+00
KR2b	編年類	2		305	6.95E-02	3.71E-02	1.87E+00	3.90E-02	1.78E+00
KR2d	別史類	2		87	1.98E-02	2.03E-02	9.76E-01	2.23E-02	8.91E-01
KR2e	雜史類	2		404	9.20E-02	3.82E-02	2.41E+00	4.00E-02	2.30E+00
KR2g	傳記類	3		56	1.28E-02	1.64E-02	7.76E-01	1.76E-02	7.27E-01
KR2i	載記類	5		128	2.92E-02	4.06E-02	7.18E-01	4.28E-02	6.81E-01
KR2j	時令類	1		1	2.28E-04	4.23E-03	5.39E-02	4.77E-03	4.77E-02
KR2k	地理類	3		51	1.16E-02	6.56E-02	1.77E-01	6.07E-02	1.92E-01
KR2m	政書類	1		2	4.56E-04	3.52E-02	1.29E-02	3.42E-02	1.33E-02

Category		Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR2o	史評類	2		120	2.73E-02	5.15E-02	5.31E-01	4.47E-02	6.12E-01
KR2p	簡帛金石類	66		47	1.07E-02	6.20E-03	1.73E+00	7.55E-03	1.42E+00
KR3	子部	0	17150		5.26E-01	8.06E-01	6.52E-01	8.37E-01	6.28E-01
KR3a	儒家類	15		912	5.32E-02	7.67E-03	6.94E+00	8.68E-03	6.13E+00
KR3b	兵家類	10		81	4.72E-03	1.74E-04	2.71E+01	2.34E-04	2.02E+01
KR3c	法家類	4		581	3.39E-02	9.63E-04	3.52E+01	1.15E-03	2.95E+01
KR3d	農家類	1		8	4.66E-04	2.94E-04	1.58E+00	3.92E-04	1.19E+00
KR3e	醫家類	9	6581	14	3.84E-01	9.15E-01	4.19E-01	9.19E-01	4.17E-01
KR3ea	醫經類	55		778	1.18E-01	5.87E-02	2.02E+00	6.06E-02	1.96E+00
KR3eb	診法類	18		11	1.68E-03	2.96E-03	5.65E-01	3.12E-03	5.36E-01
KR3ec	本草類	76		796	1.21E-01	1.13E-01	1.07E+00	1.16E-01	1.04E+00
KR3ed	醫方類	146		343	5.22E-02	1.61E-01	3.25E-01	1.67E-01	3.14E-01
KR3ee	針灸經穴類	54		161	2.45E-02	3.25E-02	7.55E-01	3.32E-02	7.38E-01
KR3ef	傷寒金匱類	85		323	4.92E-02	7.74E-02	6.36E-01	7.79E-02	6.32E-01
KR3eg	溫病類	48		102	1.55E-02	1.67E-02	9.29E-01	1.57E-02	9.88E-01
KR3eh	內科	56		203	3.09E-02	5.41E-02	5.71E-01	5.26E-02	5.88E-01
KR3ei	婦產科	73		141	2.15E-02	3.57E-02	6.02E-01	3.57E-02	6.02E-01
KR3ej	兒科	58		120	1.83E-02	3.42E-02	5.34E-01	3.42E-02	5.35E-01
KR3ek	外科	50		138	2.10E-02	3.85E-02	5.46E-01	3.78E-02	5.55E-01
KR3el	傷科	18		2	3.05E-04	1.89E-03	1.61E-01	1.80E-03	1.69E-01
KR3em	五官科	39		98	1.49E-02	1.35E-02	1.11E+00	1.25E-02	1.19E+00
KR3eo	養生類	45		212	3.23E-02	1.36E-02	2.37E+00	1.35E-02	2.39E+00
KR3ep	醫案類	103		253	3.85E-02	4.97E-02	7.75E-01	4.82E-02	8.00E-01
KR3eq	醫論醫話隨筆類	97		830	1.26E-01	3.01E-02	4.19E+00	2.85E-02	4.44E+00
KR3er	綜合醫書類	112		940	1.43E-01	1.95E-01	7.35E-01	1.93E-01	7.42E-01
KR3es	醫史考釋類	6		455	6.93E-02	8.43E-03	8.22E+00	8.33E-03	8.32E+00
KR3et	其他類	2		50	7.61E-03	2.93E-04	2.60E+01	3.41E-04	2.24E+01
KR3eu	叢書	61		611	9.30E-02	6.32E-02	1.47E+00	6.05E-02	1.54E+00
KR3f	天文算法類	16	57	5	3.32E-03	2.01E-02	1.65E-01	9.01E-03	3.69E-01

Category	Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR3fa 天文	38		19	3.65E-01	2.95E-01	1.24E+00	2.29E-01	1.60E+00
KR3fb 曆法	24		16	3.08E-01	2.04E-01	1.51E+00	1.95E-01	1.58E+00
KR3fc 算法	79		17	3.27E-01	5.01E-01	6.53E-01	5.76E-01	5.68E-01
KR3g 術數類	4		51	2.97E-03	4.77E-04	6.23E+00	6.68E-04	4.45E+00
KR3h 藝術類	4		2	1.17E-04	1.54E-05	7.59E+00	2.09E-05	5.58E+00
KR3i 譜錄類	1		1	5.83E-05	1.15E-05	5.07E+00	1.73E-05	3.37E+00
KR3j 雜家類	24		1527	8.90E-02	5.25E-03	1.70E+01	5.85E-03	1.52E+01
KR3k 類書類	22		6800	3.97E-01	4.28E-02	9.27E+00	4.65E-02	8.54E+00
KR3l 小說家類	15		549	3.20E-02	7.11E-03	4.50E+00	8.14E-03	3.93E+00
KR4 集部	0	10374		3.18E-01	1.73E-01	1.84E+00	1.43E-01	2.23E+00
KR4a 楚辭類	1		5	4.82E-04	2.54E-04	1.90E+00	2.93E-04	1.65E+00
KR4b 別集類	4		36	3.47E-03	1.74E-03	1.99E+00	2.15E-03	1.62E+00
KR4h 總集類	3		156	1.50E-02	5.23E-03	2.87E+00	7.32E-03	2.05E+00
KR4i 詩文評類	3		5	4.82E-04	7.40E-04	6.52E-01	1.06E-03	4.55E-01
KR4k 小說類	329		10172	9.81E-01	9.92E-01	9.88E-01	9.89E-01	9.91E-01
KR5 道部	0	648		1.11E-02	1.82E-03	6.12E+00	1.37E-03	8.14E+00
KR5a 洞真部	2		2	3.09E-03	1.32E-02	2.35E-01	2.27E-02	1.36E-01
KR5b 洞玄部	3		3	4.63E-03	9.53E-03	4.86E-01	1.47E-02	3.14E-01
KR5c 洞神部	11		440	6.79E-01	5.99E-01	1.13E+00	4.02E-01	1.69E+00
KR5e 太平部	1		129	1.99E-01	2.21E-01	9.02E-01	3.37E-01	5.90E-01
KR5f 太清部	1		71	1.10E-01	1.56E-01	7.02E-01	2.20E-01	4.98E-01
KR5h 續道藏	1		3	4.63E-03	1.42E-03	3.26E+00	2.98E-03	1.55E+00
KR6 佛部	0	24864		4.28E-01	3.56E-01	1.20E+00	2.98E-01	1.43E+00
KR6a 阿含部類	181		359	1.44E-02	1.44E-02	1.00E+00	2.04E-02	7.09E-01
KR6b 本緣部類	87		577	2.32E-02	1.38E-02	1.68E+00	1.78E-02	1.30E+00
KR6c 般若部類	229		517	2.08E-02	4.13E-02	5.03E-01	4.52E-02	4.60E-01
KR6d 法華部類	239		1055	4.24E-02	7.66E-02	5.54E-01	6.71E-02	6.33E-01
KR6e 華嚴部類	166		1107	4.45E-02	6.47E-02	6.88E-01	5.07E-02	8.77E-01
KR6f 寶積部類	110		233	9.37E-03	1.42E-02	6.59E-01	1.63E-02	5.75E-01

Category	Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR6g　涅槃部類	54		188	7.56E-03	1.91E-02	3.96E-01	1.83E-02	4.13E-01
KR6h　大集部類	38		249	1.00E-02	8.66E-03	1.16E+00	9.95E-03	1.01E+00
KR6i　經集部類	630		1114	4.48E-02	6.92E-02	6.48E-01	6.71E-02	6.68E-01
KR6j　密教部類	731		474	1.91E-02	6.87E-02	2.78E-01	5.56E-02	3.43E-01
KR6k　律部類	266		698	2.81E-02	7.99E-02	3.51E-01	9.07E-02	3.09E-01
KR6l　毘曇部類	40		75	3.02E-03	3.57E-02	8.46E-02	3.72E-02	8.11E-02
KR6m　中觀部類	44		82	3.30E-03	8.87E-03	3.72E-01	9.86E-03	3.34E-01
KR6n　瑜伽部類	135		129	5.19E-03	5.04E-02	1.03E-01	4.95E-02	1.05E-01
KR6o　論集部類	120		76	3.06E-03	1.67E-02	1.83E-01	1.68E-02	1.82E-01
KR6p　淨土宗部類	141		405	1.63E-02	1.71E-02	9.55E-01	1.76E-02	9.28E-01
KR6q　禪宗部類	610		11439	4.60E-01	1.76E-01	2.62E+00	1.75E-01	2.63E+00
KR6r　史傳部類	197		3566	1.43E-01	5.03E-02	2.85E+00	5.66E-02	2.53E+00
KR6s　事彙部類	145		1072	4.31E-02	4.91E-02	8.78E-01	4.48E-02	9.63E-01
KR6t　續諸宗(日本)	2		23	9.25E-04	2.30E-03	4.03E-01	1.70E-03	5.45E-01
KR6u　敦煌寫本部類	57		26	1.05E-03	9.93E-04	1.05E+00	9.96E-04	1.05E+00
KR6v　新編部類	367		1384	5.57E-02	8.61E-02	6.47E-01	8.72E-02	6.38E-01
KR6w　南傳大藏經部類	83		16	6.44E-04	3.62E-02	1.78E-02	4.36E-02	1.48E-02

2. Searching for 半夏

Total number of hits: 149565

Category	Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR0　四部	0	149410		9.99E-01	6.42E-01	1.56E+00	7.00E-01	1.43E+00
KR1　經部	0	2		1.34E-05	6.02E-03	2.22E-03	4.58E-03	2.92E-03
KR1d　禮類	4		1	5.00E-01	1.16E-01	4.32E+00	2.02E-01	2.47E+00
KR1j　小學類	6		1	5.00E-01	6.06E-01	8.26E-01	2.98E-01	1.68E+00
KR2　史部	0	6		4.02E-05	1.49E-02	2.70E-03	1.60E-02	2.51E-03
KR2a　正史類	6		3	5.00E-01	6.85E-01	7.30E-01	6.86E-01	7.28E-01
KR2d　別史類	2		1	1.67E-01	2.03E-02	8.21E+00	2.23E-02	7.49E+00

Category	Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR2j 時令類	1		2	3.33E-01	4.23E-03	7.88E+01	4.77E-03	6.98E+01
KR3 子部	0	149376		1.00E+00	8.06E-01	1.24E+00	8.37E-01	1.19E+00
KR3e 醫家類	9	149331	137	1.00E+00	9.15E-01	1.09E+00	9.19E-01	1.09E+00
KR3ea 醫經類	55		735	4.93E-03	5.87E-02	8.39E-02	6.06E-02	8.13E-02
KR3eb 診法類	18		23	1.54E-04	2.96E-03	5.20E-02	3.12E-03	4.93E-02
KR3ec 本草類	76		5767	3.87E-02	1.13E-01	3.42E-01	1.16E-01	3.33E-01
KR3ed 醫方類	146		32559	2.18E-01	1.61E-01	1.36E+00	1.67E-01	1.31E+00
KR3ee 針灸經穴類	54		697	4.67E-03	3.25E-02	1.44E-01	3.32E-02	1.41E-01
KR3ef 傷寒金匱類	85		12524	8.39E-02	7.74E-02	1.09E+00	7.79E-02	1.08E+00
KR3eg 溫病類	48		1977	1.33E-02	1.67E-02	7.92E-01	1.57E-02	8.43E-01
KR3eh 內科	56		12155	8.15E-02	5.41E-02	1.51E+00	5.26E-02	1.55E+00
KR3ei 婦產科	73		6148	4.12E-02	3.57E-02	1.15E+00	3.57E-02	1.16E+00
KR3ej 兒科	58		4945	3.31E-02	3.42E-02	9.69E-01	3.42E-02	9.70E-01
KR3ek 外科	50		3114	2.09E-02	3.85E-02	5.42E-01	3.78E-02	5.52E-01
KR3el 傷科	18		319	2.14E-03	1.89E-03	1.13E+00	1.80E-03	1.18E+00
KR3em 五官科	39		1133	7.59E-03	1.35E-02	5.64E-01	1.25E-02	6.06E-01
KR3eo 養生類	45		1577	1.06E-02	1.36E-02	7.77E-01	1.35E-02	7.83E-01
KR3ep 醫案類	103		14749	9.89E-02	4.97E-02	1.99E+00	4.82E-02	2.05E+00
KR3eq 醫論醫話隨筆類	97		3266	2.19E-02	3.01E-02	7.26E-01	2.85E-02	7.69E-01
KR3er 綜合醫書類	112		37497	2.51E-01	1.95E-01	1.29E+00	1.93E-01	1.30E+00
KR3es 醫史考釋類	6		418	2.80E-03	8.43E-03	3.33E-01	8.33E-03	3.36E-01
KR3et 其他類	2		4	2.68E-05	2.93E-04	9.15E-02	3.41E-04	7.87E-02
KR3eu 叢書	61		9587	6.43E-02	6.32E-02	1.02E+00	6.05E-02	1.06E+00
KR3f 天文算法類	16	12		8.03E-05	2.01E-02	3.99E-03	9.01E-03	8.92E-03
KR3fa 天文	38		5	4.17E-01	2.95E-01	1.41E+00	2.29E-01	1.82E+00
KR3fb 曆法	24		5	4.17E-01	2.04E-01	2.04E+00	1.95E-01	2.14E+00
KR3fc 算法	79		2	1.67E-01	5.01E-01	3.33E-01	5.76E-01	2.89E-01
KR3j 雜家類	24		8	5.36E-05	5.25E-03	1.02E-02	5.85E-03	9.16E-03
KR3k 類書類	22		23	1.54E-04	4.28E-02	3.60E-03	4.65E-02	3.31E-03

Category		Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR3l	小說家類	15		2	1.34E-05	7.11E-03	1.88E-03	8.14E-03	1.65E-03
KR4	集部	0	26		1.74E-04	1.73E-01	1.01E-03	1.43E-01	1.22E-03
KR4k	小說類	329		26	1.00E+00	9.92E-01	1.01E+00	9.89E-01	1.01E+00
KR6	佛部	0	116		7.76E-04	3.56E-01	2.18E-03	2.98E-01	2.60E-03
KR6d	法華部類	239		1	8.62E-03	7.66E-02	1.13E-01	6.71E-02	1.29E-01
KR6e	華嚴部類	166		1	8.62E-03	6.47E-02	1.33E-01	5.07E-02	1.70E-01
KR6k	律部類	266		24	2.07E-01	7.99E-02	2.59E+00	9.07E-02	2.28E+00
KR6p	淨土宗部類	141		3	2.59E-02	1.71E-02	1.52E+00	1.76E-02	1.47E+00
KR6q	禪宗部類	610		68	5.86E-01	1.76E-01	3.33E+00	1.75E-01	3.35E+00
KR6r	史傳部類	197		11	9.48E-02	5.03E-02	1.89E+00	5.66E-02	1.68E+00
KR6s	事彙部類	145		3	2.59E-02	4.91E-02	5.26E-01	4.48E-02	5.78E-01
KR6v	新編部類	367		5	4.31E-02	8.61E-02	5.01E-01	8.72E-02	4.94E-01

3. Searching for 國語

Total number of hits: 3060

Category		Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR0	四部	0	1078		3.72E-01	6.42E-01	5.80E-01	7.00E-01	5.32E-01
KR1	經部	0	88		8.16E-02	6.02E-03	1.36E+01	4.58E-03	1.78E+01
KR1j	小學類	6		88	1.00E+00	6.06E-01	1.65E+00	2.98E-01	3.35E+00
KR2	史部	0	62		5.75E-02	1.49E-02	3.86E+00	1.60E-02	3.60E+00
KR2a	正史類	6		13	2.10E-01	6.85E-01	3.06E-01	6.86E-01	3.05E-01
KR2b	編年類	2		2	3.23E-02	3.71E-02	8.69E-01	3.90E-02	8.26E-01
KR2j	時令類	1		1	1.61E-02	4.23E-03	3.81E+00	4.77E-03	3.38E+00
KR2k	地理類	3		19	3.06E-01	6.56E-02	4.67E+00	6.07E-02	5.05E+00
KR2m	政書類	1		1	1.61E-02	3.52E-02	4.58E-01	3.42E-02	4.71E-01
KR2o	史評類	2		26	4.19E-01	5.15E-02	8.14E+00	4.47E-02	9.39E+00
KR3	子部	0	880		8.16E-01	8.06E-01	1.01E+00	8.37E-01	9.76E-01
KR3a	儒家類	15		3	3.41E-03	7.67E-03	4.45E-01	8.68E-03	3.93E-01
KR3e	醫家類	9	154		1.75E-01	9.15E-01	1.91E-01	9.19E-01	1.90E-01

Category		Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR3ea	醫經類	55		36	2.34E-01	5.87E-02	3.98E+00	6.06E-02	3.86E+00
KR3ec	本草類	76		60	3.90E-01	1.13E-01	3.45E+00	1.16E-01	3.35E+00
KR3ed	醫方類	146		1	6.49E-03	1.61E-01	4.04E-02	1.67E-01	3.90E-02
KR3ee	針灸經穴類	54		4	2.60E-02	3.25E-02	7.99E-01	3.32E-02	7.82E-01
KR3ef	傷寒金匱類	85		8	5.19E-02	7.74E-02	6.71E-01	7.79E-02	6.67E-01
KR3eo	養生類	45		18	1.17E-01	1.36E-02	8.60E+00	1.35E-02	8.65E+00
KR3eq	醫論醫話隨筆類	97		11	7.14E-02	3.01E-02	2.37E+00	2.85E-02	2.51E+00
KR3es	醫史考釋類	6		4	2.60E-02	8.43E-03	3.08E+00	8.33E-03	3.12E+00
KR3eu	叢書	61		12	7.79E-02	6.32E-02	1.23E+00	6.05E-02	1.29E+00
KR3f	天文算法類	16	8		9.09E-03	2.01E-02	4.52E-01	9.01E-03	1.01E+00
KR3fa	天文	38		5	6.25E-01	2.95E-01	2.12E+00	2.29E-01	2.73E+00
KR3fb	曆法	24		2	2.50E-01	2.04E-01	1.23E+00	1.95E-01	1.28E+00
KR3fc	算法	79		1	1.25E-01	5.01E-01	2.50E-01	5.76E-01	2.17E-01
KR3j	雜家類	24		19	2.16E-02	5.25E-03	4.11E+00	5.85E-03	3.69E+00
KR3k	類書類	22		692	7.86E-01	4.28E-02	1.84E+01	4.65E-02	1.69E+01
KR3l	小說家類	15		4	4.55E-03	7.11E-03	6.40E-01	8.14E-03	5.58E-01
KR4	集部	0	48		4.45E-02	1.73E-01	2.58E-01	1.43E-01	3.12E-01
KR4b	別集類	4		1	2.08E-02	1.74E-03	1.19E+01	2.15E-03	9.70E+00
KR4h	總集類	3		1	2.08E-02	5.23E-03	3.98E+00	7.32E-03	2.85E+00
KR4k	小説類	329		46	9.58E-01	9.92E-01	9.66E-01	9.89E-01	9.69E-01
KR5	道部	0	1		3.45E-04	1.82E-03	1.90E-01	1.37E-03	2.52E-01
KR5c	洞神部	11		1	1.00E+00	5.99E-01	1.67E+00	4.02E-01	2.48E+00
KR6	佛部	0	1816		6.27E-01	3.56E-01	1.76E+00	2.98E-01	2.10E+00
KR6b	本緣部類	87		4	2.20E-03	1.38E-02	1.59E-01	1.78E-02	1.23E-01
KR6c	般若部類	229		18	9.91E-03	4.13E-02	2.40E-01	4.52E-02	2.19E-01
KR6d	法華部類	239		41	2.26E-02	7.66E-02	2.95E-01	6.71E-02	3.37E-01
KR6e	華嚴部類	166		13	7.16E-03	6.47E-02	1.11E-01	5.07E-02	1.41E-01
KR6f	寶積部類	110		10	5.51E-03	1.42E-02	3.88E-01	1.63E-02	3.38E-01
KR6g	涅槃部類	54		14	7.71E-03	1.91E-02	4.04E-01	1.83E-02	4.21E-01

Category		Texts	Sum	Matches	MF	ICF	MF/ICF	ISF	MF / ISF
KR6i	經集部類	630		24	1.32E-02	6.92E-02	1.91E-01	6.71E-02	1.97E-01
KR6j	密教部類	731		14	7.71E-03	6.87E-02	1.12E-01	5.56E-02	1.39E-01
KR6k	律部類	266		59	3.25E-02	7.99E-02	4.07E-01	9.07E-02	3.58E-01
KR6l	毘曇部類	40		7	3.85E-03	3.57E-02	1.08E-01	3.72E-02	1.04E-01
KR6m	中觀部類	44		5	2.75E-03	8.87E-03	3.10E-01	9.86E-03	2.79E-01
KR6n	瑜伽部類	135		26	1.43E-02	5.04E-02	2.84E-01	4.95E-02	2.89E-01
KR6p	淨土宗部類	141		5	2.75E-03	1.71E-02	1.61E-01	1.76E-02	1.57E-01
KR6q	禪宗部類	610		13	7.16E-03	1.76E-01	4.07E-02	1.75E-01	4.09E-02
KR6r	史傳部類	197		57	3.14E-02	5.03E-02	6.24E-01	5.66E-02	5.54E-01
KR6s	事彙部類	145		1453	8.00E-01	4.91E-02	1.63E+01	4.48E-02	1.79E+01
KR6u	敦煌寫本部類	57		1	5.51E-04	9.93E-04	5.54E-01	9.96E-04	5.53E-01
KR6v	新編部類	367		48	2.64E-02	8.61E-02	3.07E-01	8.72E-02	3.03E-01
KR6w	南傳大藏經部類	83		4	2.20E-03	3.62E-02	6.09E-02	4.36E-02	5.06E-02