Library Hi Tech

THC-DAT: a document analysis tool based on topic hierarchy and context information

Jing Chen Tian Tian Wang Quan Lu

Article information:

To cite this document:
Jing Chen Tian Tian Wang Quan Lu , (2016),"THC-DAT: a document analysis tool based on topic hierarchy and context information", Library Hi Tech, Vol. 34 Iss 1 pp. 64 - 86
Permanent link to this document:
http://dx.doi.org/10.1108/LHT-07-2015-0074

Users who downloaded this article also downloaded:

For Authors

About Emerald www.emeraldinsight.com

Received 15 July 2015
Revised 23 September 2015
Accepted 1 October 2015

THC-DAT: a document analysis tool based on topic hierarchy and context information

Jing Chen and Tian Tian Wang
School of Information Management, Central China Normal University, Wuhan, PR China, and
Quan Lu
Center for Studies of Information Resources, Wuhan University, Wuhan, PR China

Abstract

Purpose – The purpose of this paper is to propose a novel within-document analysis tool (DAT) topic hierarchy and context-based document analysis tool (THC-DAT) which enables users to interactively analyze any multi-topic document based on fine-grained and hierarchical topics automatically extracted from it. THC-DAT used hierarchical latent Dirichlet allocation method and took the context information into account so that it can reveal the relationships between latent topics and related texts in a document.
Design/methodology/approach – The methodology is a case study. The authors reviewed the related literature first, then utilized a general "build and test" research model. After explaining the model, interface and functions of THC-DAT, a case study was presented using a scholarly paper that was analyzed with the tool.
Findings – THC-DAT can organize and serve document topics and texts hierarchically and context based, which overcomes the drawbacks of traditional DATs. The navigation, browse, search and comparison functions of THC-DAT enable users to read, search and analyze multi-topic document efficiently and effectively.
Practical implications – It can improve the document organization and services in digital libraries or e-readers, by helping users to interactively read, search and analyze documents efficiently and effectively, exploringly learn about unfamiliar topics with little cognitive burden, or deepen their understanding of a document.
Originality/value – This paper designs a tool THC-DAT to analyze document in a THC way. It contributes to overcoming the coarse-analysis drawbacks of existing within-DATs.

Keywords Digital libraries, E-readers, Document analysis, Context information, hLDA, Multi-topic documents
Paper type Technical paper

1. Introduction

With the growing availability of electronic documents, there are more and more multi-topic documents. Multi-topic documents arise in various application domains including scientific articles, news stories, patents, judgments and decisions reported in courts and tribunals (case law documents), and speeches delivered by plenary sessions (e.g. parliamentary debates) (Andrea and George, 2013). The common characteristic of these documents is that they may discuss various topics that are related to the article's main topic. For instance, scientific articles in the field of information science usually

The authors gratefully acknowledge the financial support for this work provided by National Natural Science Foundation of China (No:71303089, 71273195 and 71420107026) and the National Basic Research Program of China (973 Program, No: 904171200).

© Emerald Group Publishing Limited
0737-8831
DOI 10.1108/LHT-07-2015-0074

involve principles and techniques from informatics and library science, computer science, cognitive science and statistics.

Document analysis and topic extraction tools can help researchers to organize, extract and interpret knowledge efficiently and provide new ideas for their scientific research domains. However, predominant document analysis tools (DATs), including within-DATs, mainly focus on extracting specific features from full text aspect in the past years. These tools allow users to retrieve from titles, paragraphs, even the full text, only in a coarse-grained perspective. Generally speaking, they analyze the words in document by either counting them or indexing their places of occurrence only in full text, such as Concordance (Watt, 2015), FeatureLens (Don *et al.*, 2007), iSee (Sun *et al.*, 2005), TextArc (Paley, 2002) and Jigsaw (Stasko *et al.*, 2008), and so on. Besides, they also have some other flaws, either ignoring the context information, or considering document as a smooth liner structure (Du *et al.*, 2012). So far, existing tools provide poor document analysis capabilities because of the above flaws.

This paper proposed a new multi-topic DAT topic hierarchy and context-based document analysis tool (THC-DAT) based on THC information, which enables users to search, browse and analyze within document more efficiently and effectively. In our research, a document is looked as a collection of text segments, and each paragraph is one segment. THC-DAT visualizes a topic hierarchy tree, in which users can search and browse according to their interested topics to obtain relevant paragraphs and analyze hierarchical structure of the document. Furthermore, by context information, a user can quickly grasp the distribution of topics in the paper so that (s)he can find out the relationships between paragraphs and between topics. Generally, THC-DAT has the following characteristics:

(1) Topic oriented: it extracts topics with hierarchical structure, and the topic term is abstraction of the corresponding text. So, by viewing the topics, user can have a preliminary understanding of the corresponding texts and their relationship.

(2) Multi-grained: it organizes the document as a hierarchy tree, and text more general will correspond to the root while text more specialized will correspond to the leaf. So it can help user analysis text from abstraction to concretion according to the hierarchical structure.

(3) Semantic aggregated: by merging adjacent paragraphs based on context information, the tool can provide more accurate analysis results to meet user's requirement.

The rest of the paper is arranged as follows. Section 2 reviews related researches. Section 3 describes the model of THC-DAT. Section 4 introduces the interface and functions of THC-DAT. Section 5 shows a case study of THC-DAT and the discussion and Section 6 presents some concluding remarks.

2. Related researches

With the rapid increase of electronic document resources in recent years, there is a large amount of work devoted to document retrieval and analysis. Tools which help users to analyze within document are becoming a research hotspot, and there are some considerable and valuable works on within-document retrieval and analysis.

TileBars (Hearst, 1995) is an early and influential retrieval tool for users to search within document, which takes a set of search terms and creates a matrix of titles, where every line represents an entire text and each column is on behalf of a block of text in the document. TileBars visualizes query term frequency and term location for users who want to identify whether a document was relevant to the given query terms.

Scrollbar (Byrd, 1999) and TextArc (Paley, 2002) were constructed by TileBars-like ideas. These two tools stated the location of query terms within a document and placed small icons of corresponding colored and highlighted query terms, helping users quickly find relevant sections. Overview scrollbar (Ko Mizoguchi, 2013) was designed on the basis of scrollbar, and it displayed an overview of the entire document and implemented four types of overview scrollbars that used different compression methods to render the overviews. Their functions were similar to THC-DAT, but did not include hierarchy and context.

SCAN (Whittaker *et al.*, 1999) provided a straightforward histogram to state the relevant sections and granular occurrences of each term. Besides, ProfileSkim (Harper *et al.*, 2004) divided a document into multiple tiles, each tile had multiple text window with fixed limited size. However, the text window and the division of tile did not stress on logical relationship. Xed (Hadjar *et al.*, 2004) considered the physical and logical structure for PDF documents, and it mainly focussed on automatically extracting document's layout structures for automatic analysis. SmartSkim (Harper *et al.*, 2002) was a prototype of ProfileSkim and very similar to TileBars and SCAN.

ThemeRiver (Havre *et al.*, 2002) visually depicted the thematic variations over time within a document. It used a river metaphor to illustrate but only the thematic strength. Rivers flowing from left to right represented the time series, and the theme of the text was displayed by different colors of ribbon. Tiara (Liu *et al.*, 2012) applied latent Dirichlet allocation (LDA)-based topic analysis to automatically derive topics and content evolution over time. It could not only illustrate thematic strength variations over time, but also depict the detail content in keywords and visualize complex text summarization results. DocuBurst (Collins *et al.*, 2009) was designed to depict document structure through radial, space-filling layout of hyponymy, overlaid with occurrence counts of words to provide visual summaries.

The common drawbacks of these tools are that ignoring the context and the latent hierarchical structures in document. Traditional analysis ideas often consider a document as a sequential structure (Du *et al.*, 2012). However, many documents, especially scientific research papers, may come naturally with complex hierarchical structure, and paragraphs in different locations may have same topic. Each document can be seen as a system (Jiang *et al.*, 1983), which contains a trunk, branches and leaves. They constitute an organic whole, and associate with each other. Research indicated that hierarchical structure analysis could reveal the structure better from a macro view to grasp the full document than linear (Zhang *et al.*, 2006; Salton and Allan, 1994).

Context also plays an important role in understanding full text (Messelodi and Modena, 1996; Donoser *et al.*, 2010). In recent years, the study of context has been gaining popularity. Relevant researches have given different definitions. Crystal (1991) considers it indicating the total linguistic and non-linguistic background of a text. Brown and Yule (1983) believed that context is the physical environment where a word is used, while Dey (2005) considered it as "any information that can be used to characterize the situation of an entity." Based on the review of existing definitions on context, we conclude in this research that context is the specific language environment that helps to explain relationships between paragraphs.

Context information bridges the relationship between sections, reveals underlying background knowledge and helps researchers locate interesting or valuable information. It is an important approach for document analysis and can provide a higher satisfaction for users (Brown and Jones, 2001; Moskovitch *et al.*, 2007). Existing researches about context generally focussed on eliminating ambiguity (Agirre *et al.*, 2014; Brosseau-Villeneuve *et al.*, 2014). They considered context as a specific semantic environment where words occur, and by identifying semantic relationships about words, typically based on WordNet, users could understand the intended meaning of a word in the given context. In addition, some researchers also used context to study other fields, for example, when trying to improve the information retrieval effect (Tanveer and Tiwary, 2005), context was considered to indicate some semantic gap or background within the document, and the query word's location was considered in order to provide more suitable and accurate retrieval results for users. When studying automatic summarization, Suo *et al.* (2007) took context as a form of knowledge leading a close link among paragraphs, and combined word's repetition and paragraph's distance in context to determine the boundary for multiple topic partition, which improves the topic coverage and abstract extraction accuracy greatly.

Based on Suo's research, the paragraph distance is taken into consideration and the paragraph ordinal is identified as context information for our tool. the paragraph ordinal represents a specific semantic background, and is a link throughout the full text to explain some relationship among paragraphs. It would produce better segmentation effects to calculate the distance among paragraphs under the same topic and merge paragraphs to construct semantic units. By building and utilizing the semantic units, THC-DAT may provide more accurate and dynamic results and perfect user experience, so user can browse and analyze document contents, and study the document by comparing topics and semantic units, conveniently and interactively.

3. Model of THC-DAT

The proposed model of THC-DAT is described systematically in this section. It is depicted in Figure 1, containing three major modules presented as follows: document preprocessing, topic hierarchy generation and paragraphs merging based on context information.

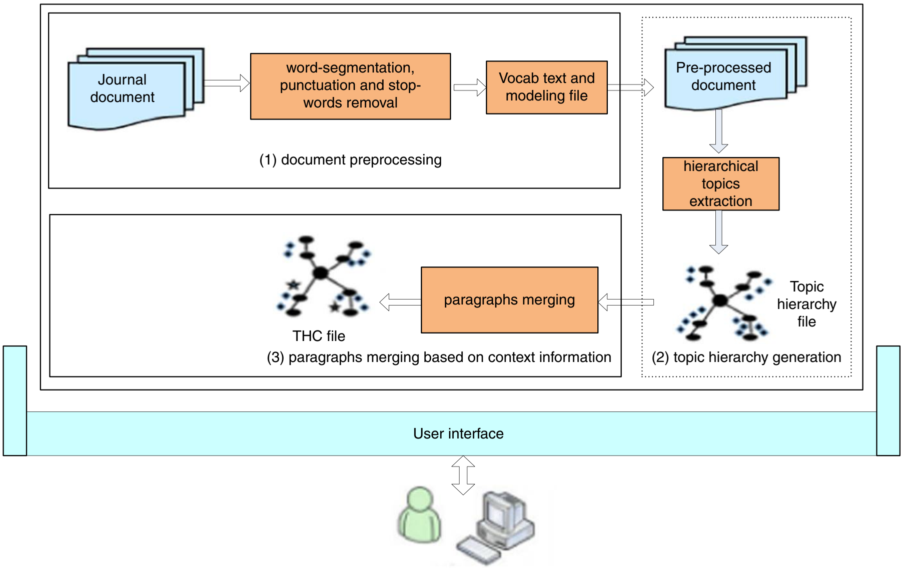

Journal document
word-segmentation, punctuation and stop-words removal
Vocab text and modeling file
(1) document preprocessing
Pre-processed document
hierarchical topics extraction
Topic hierarchy file
(2) topic hierarchy generation
paragraphs merging
THC file
(3) paragraphs merging based on context information
User interface

Figure 1.
Model of THC-DAT

3.1 Document Preprocessing

First, we partition a document into many segments, each paragraph will be regarded as one segment, and then we operate word-segmentation, remove stop-words and punctuations. So the non-repeated vocab text of the document which consists of a list of terms and term's sequence number is generated. Each line in the vocab text is composed of a specific term's serial number and the corresponding term. Then, the data file which will be used in module 2 is modeled, in accordance with input format specification of the hierarchical latent Dirichlet allocation (hLDA) package developed by Blei (2009), as shown in Figure 2.

Each line in the modeling file in Figure 2 represents one paragraph, where [Mi] is the number of unique terms in the paragraph, followed closely by each term's serial number in the vocab text, and [count] associated with each term is how many times the term appears in the paragraph.

The vocab text and the modeling file are composed of the pre-processed file which will be input to module 2 to extract the topic for the document to be analyzed.

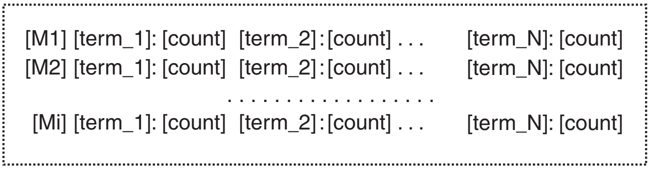
3.2 Topic hierarchy generation

In the module 2, the pre-processed file is as the input file and after hierarchical topics extraction, we get the topic hierarchy file. Intuitively, the hierarchical topics extraction method we used here is the topic modeling method.

Topic model is an efficient method that can help researchers to explore the underlying meaningful information (Misra *et al.*, 2011; Vo and Ock, 2015). hLDA model (Blei *et al.*, 2003b) is a very successful topic model for document analysis and informational retrieval (Venkatesh, 2013). It is based on LDA which allows users to analyze of corpus, and extract the topics that combined to form its documents (Blei *et al.*, 2003a). Furthermore, hLDA can organize the topics into a hierarchy tree with fixed depth (Blei *et al.*, 2003b). Here, a topic is defined to be a probability distribution across words from a vocabulary. The characteristic of hLDA is that no specific knowledge of the topics of the document(s) or a preset structure of the tree is needed to infer a hierarchy from data. The structure of the tree depends on the document data and the parameters. When analyzing a document collection with hLDA to mine a topic hierarchy, users can adjust various parameters to get an ideal tree that meets their needs (Blei *et al.*, 2010). For the advantages of hLDA mentioned above, the hierarchical topics extraction method used here is hLDA.

Figure 3 shows the structure of a topic hierarchy tree of the pre-processed document which is generated by hLDA.

As shown in Figure 3, for the hierarchy tree, each node in the tree represents a topic with its terms and has a set of corresponding paragraphs. Topic 0 is the root topic, and (Topic *a*, Topic *b*, …, Topic *m*) are all sub-topics of Topic 0, in the same way, Topic *n* is also a sub-topic of Topic *a*. A path from the root node to a leaf node, such as the path from Topic 0 to Topic *n,* represents the topics evolve from more abstract to more concrete, the corresponding texts become more concrete too. The corresponding paragraphs of a higher level topic node cover the corresponding passages of a lower level topic node if they are in



```
[M1] [term_1]: [count]  [term_2]:[count] . . .     [term_N]: [count]
[M2] [term_1]: [count]  [term_2]:[count] . . .     [term_N]: [count]
                        . . . . . . . . . . . . . . . . . .
 [Mi] [term_1]: [count]  [term_2]:[count] . . .     [term_N]: [count]
```

Figure 2.
Modeling file format

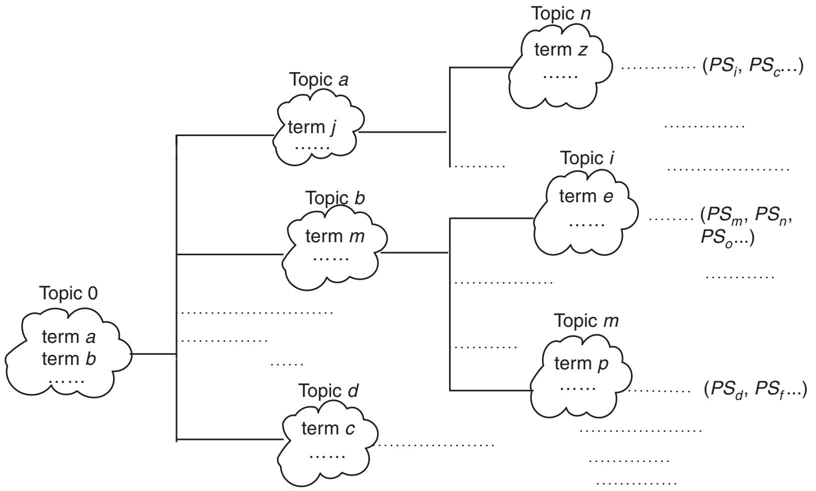


Figure 3. Structure of topic hierarchy file

the same path. Given an example, Topic 0 is corresponding to the full text and covers the corresponding paragraphs of Topic *a*, Topic *b* and so on. Besides, each paragraph belongs to one path in the tree, and each path can be shared by many paragraphs having the same topic. For example, the formula (PS_i, PS_c, …) in Figure 3 represents the paragraph set which belongs to the topic path from Topic 0 through Topic *a* to topic *n* and so on. Each topic is a distribution of terms in vocab text, a high-ranking term in a topic has higher probability than its later ones so it is better to represent the topic.

3.3 Paragraphs merging based on context information

As mentioned above, paragraphs assigned to the same path have the same topic and can be treated as one semantic unit. According to existing linguistic researches, paragraph can be thought as a grammatical unit in the following sense: a functional structure assigned to sentences (van Dijk, 1983). And paragraphs with closer space distance typically have more coherent relation and anaphora, and have less topic shift (Zadrozny and Jensen, 1991). Furthermore, in a multiple-topics document, the probability for paragraphs with long space distance belonging to a same topic is very low (Suo *et al.*, 2007). So we thought based on topic hierarchy mining, combining THC could better reveal the semantic units and their relationships in documents, and proposed a distance-based method to merge paragraphs in the same topic based on context location information. By organizing and displaying semantic units and their relationships in a document, it could help users to understand and analyze the document more effectively.

The space distance between two paragraphs can be calculated by the following formula:

$$D(PS_i, PS_j) = \left|\mathrm{Loc}(PS_i) - \mathrm{Loc}(PS_j)\right| \tag{1}$$

where n is the total number of paragraphs in a document, PS_i, PS_j ($i = 1,2, \ldots, n$, $j = 1,2, \ldots, n$) are two paragraphs, and Loc(PS_i) is the location information of paragraph PS_i and is represented by the paragraph's ordinal in the document.

Then, we used threshold method to merge relative paragraphs belonging to the same topic path according to the above context space distance. The merging result depends on the different space distance threshold. If the distance between paragraphs is more than a preset threshold T^*, the paragraphs will not be merged. On the contrary, paragraphs having distance within threshold value will be merged. The merging process is presented as follows:

(1) for the paragraphs sharing the same topic path, rank them according to the space location information and put them in numerical order.

(2) calculate the space distance D of arbitrary two paragraphs by Formula (1).

(3) compare D with threshold T^*:

- if $D > T^*$, not merge; and
- if $D \leqslant T^*$, the paragraphs within the threshold will be merged to one semantic unit.

Then we studied the method to choose an ideal threshold. In our model, the merging frequency of the paragraphs in all paths is used to determine the ideal threshold T^*, as shown in Formula (2). It is reasonable that T^* is an integer and should be greater than or equal to 0:

$$T^* = \arg\min_D F(D) \tag{2}$$

where $F(D)$ denotes the paragraph merging frequencies within the distance D for all paths and can be calculated by the following formula:

$$F(D) = \sum_{P_1}^{P_n} f(D) \tag{3}$$

here, P_n represents the path n, and $f(D)$ is the paragraph merging frequencies within the distance D for each path.

Obviously, the lower paragraphs merging frequency means the fewer semantic generality and better semantics accuracy in merging paragraphs. It also means more fine-grained within-document analysis based on THC information. So the ideal T^* for most fine-grained within-document analysis is the distance threshold which makes the paragraphs merging frequency minimum. So this model could support multi-grained within-document analysis by topic's different hierarchy and various T^*.

So, based on the topic hierarchy file shown in Figure 3, the paragraphs can be merged based on threshold T^*, and THC-DAT built a THC file through paragraphs merging. The structure of the THC file is shown in Figure 4.

As Figure 4 depicts, paragraphs in the same path are merged into semantic units in each path. For example, the formula (PS_d, PS_f, …) in Figure 4 represents the semantic unit set belonging to the path from Topic 0 to Topic m, and (PS_d, PS_f) represents a semantic unit in the set, which indicates that paragraph d and paragraph f are merged into a semantic unit because they shares the same path and are close in space distance.

4. Interface and functions of THC-DAT

As described above, THC-DAT is designed to help users browse and analyze a document in a multi-grained, topic oriented and context-based way. The main functions of THC-DAT include:

- Partition the document hierarchically into unit sets and then annotate a topic for each unit set to build the topic hierarchy and visualize it as a topic hierarchy tree. THC-DAT helps users to quickly understand topics of the document from coarse-grained to fine-grained by viewing the topic hierarchy tree.

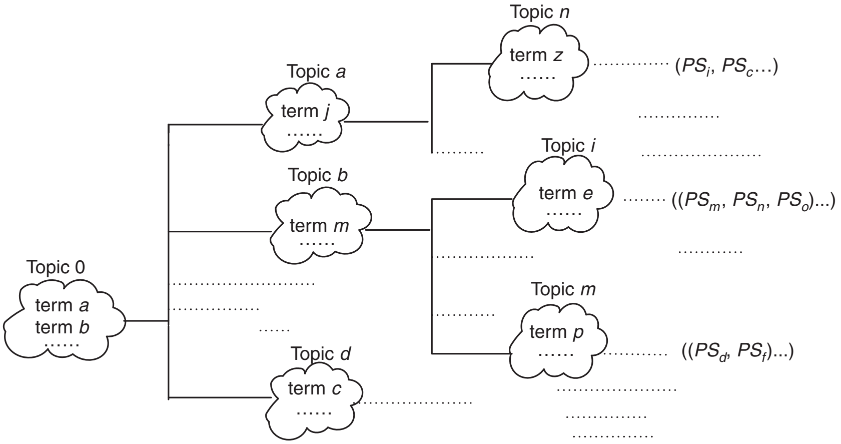


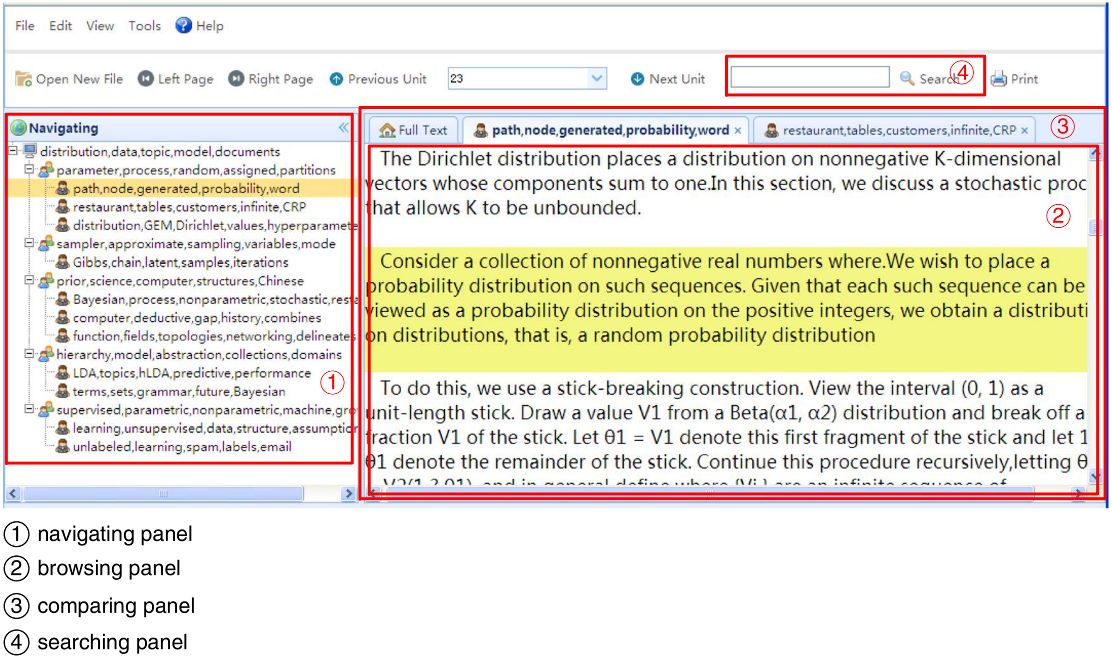
Figure 4.
Structure of THC file

- Show the corresponding semantic units for every topic in the topic hierarchy tree. It is convenient to carry out joint reading of the topics and their corresponding texts. Users can browse and compare the related semantic units for every topic in the topic hierarchy tree, so, (s)he can deeply learn the difference and connection of inter-topic and intra-topic, and do some accurate reading.
- Search query term not only in the topic hierarchy tree but also in the full text of the document. When user inputs a query term, the related topic node(s) in the topic hierarchy tree and the matched term in the document will be highlighted, so users can get a clear idea of the importance of the query term and its distribution in the full text.

The authors implemented the model we proposed in THC-DAT. It is a JAVA-based system. The interface of THC-DAT is shown in Figure 5. There are four panels in the interface: navigating panel, browsing panel, comparing panel and searching panel:

(1) Navigating panel: in the navigating panel, users can observe and operate the topic hierarchy tree of a document, which has one root node and several lower level nodes and each node is represented by top five terms. The depth of the tree depends on THC-DAT's parameter setting. The topic hierarchy tree shows all the topics in the hierarchy. By viewing the tree, users will have a primary understanding of the relations between topics quickly, and by clicking any topic node in the tree, user can browse the corresponding semantic units automatically highlighted by THC-DAT. As mentioned above, each topic node corresponds with a series of semantic units, so the highlighted semantic units with different location in the original document can share the same topic. Meanwhile, the higher level topic is more general than lower level ones in a path. Hence, the semantic units related to a parent topic node are union set of the semantic units related to all its children nodes. So users could also grasp the distribution of topics in the full text quickly by clicking nodes in the topic hierarchy tree.

(2) Browsing panel: when a topic node in the navigation panel is clicked, the corresponding semantic units will be highlighted in full text page in the browsing panel synchronously. First, as we mentioned above, the higher level topic is more abstract than the lower level ones, so users can browse and compare the text from



distribution,data,topic,model,documents
parameter,process,random,assigned,partitions
path,node,generated,probability,word
restaurant,tables,customers,infinite,CRP
distribution,GEM,Dirichlet,values,hyperparamete
sampler,approximate,sampling,variables,mode
Gibbs,chain,latent,samples,iterations
prior,science,computer,structures,Chinese
Bayesian,process,nonparametric,stochastic,resta
computer,deductive,gap,history,combines
function,fields,topologies,networking,delineates
hierarchy,model,abstraction,collections,domains
LDA,topics,hLDA,predictive,performance
terms,sets,grammar,future,Bayesian
supervised,parametric,nonparametric,machine,gro
learning,unsupervised,data,structure,assumptio
unlabeled,learning,spam,labels,email

The Dirichlet distribution places a distribution on nonnegative K-dimensional vectors whose components sum to one.In this section, we discuss a stochastic proc that allows K to be unbounded.

Consider a collection of nonnegative real numbers where.We wish to place a probability distribution on such sequences. Given that each such sequence can be viewed as a probability distribution on the positive integers, we obtain a distributi on distributions, that is, a random probability distribution

To do this, we use a stick-breaking construction. View the interval (0, 1) as a unit-length stick. Draw a value V1 from a Beta(α1, α2) distribution and break off a fraction V1 of the stick. Let θ1 = V1 denote this first fragment of the stick and let 1 θ1 denote the remainder of the stick. Continue this procedure recursively,letting θ

Figure 5.
Interface and panels of THC-DAT

abstraction to concretion by clicking nodes along a path. Then, THC-DAT supports users to browse through the semantic units belonging to the same topic rapidly with their context information. By clicking the previous unit button or next unit button, or selecting any unit from the drop-drown list box directly, users can shift to any semantic unit belonging to the same topic quickly. So users could have understanding of a topic's content and its distribution in the whole document rapidly and also have a clear idea of the context. For example, if a topic has one semantic unit appearing in the literature review section to introduce existing researches related to a method, and another semantic unit appearing in the method section to describe the use of this method in this document, users will quickly understand that these two units are closely related, the former is the basis of the latter, and this method plays a basic role in the paper.

(3) Comparing panel: in THC-DAT, every topic having been clicked in the navigating panel owns an independent tab page in the comparing panel. To be clear, the navigation panel and the browsing panel already enable users to compare between topics, but existing researches had shown that the information-seeking and content-viewing that were not on the same page would greatly enhance the user experience (Marchionini and Komlodi, 1998; Ahmed *et al.*, 2006; Bates, 2002; Davis and Price, 2006; Berg *et al.*, 2010), so the comparing panel will help users analyze and compare between topics more conveniently and efficiently. Typical comparing scenarios may be comparing topics in a path and comparing topics not in the same path but at the same level.

(4) Searching panel: users can also input any query term in the searching panel to search within both the topic hierarchy and the full text. The query term will be highlighted at every matching location in the full text. However, if it appears in the topic hierarchy, the term will be highlighted in every matched topic node also. By searching term in the full text, topic or logical units in a topic, THC-DAT allows users locate the position of the query term quickly. Further, THC-DAT helps users to find out the relationship between the query term and other topic terms roughly and browse the corresponding semantic units if the query term appears in the topic tree. THC-DAT also helps users observe the term's distribution in the document or in topics. The search function provides a way to help users quickly analyze terms in multiple-topics document during the preliminary study, so it is very valuable and effective.

In general, through above functions and their combinations, THC-DAT can organize and serve document topics and texts in a hierarchical topics and context way, which overcomes the drawbacks of traditional DATs and will enable users to read, search and analyze document more efficiently and effectively.

5. Case study and discussion

5.1 Experimental data-sets

We selected the case paper which is titled "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies." It is published in the January 2010 issue (Volume 57, No. 2) of *Journal of the Association for Computing Machinery*. It discussed how the nested Chinese restaurant process (nCRP) which is one kind of stochastic process (by simulating the probability of any customer belonging to any table) used in a Bayesian nonparametric statistics to build topic hierarchy tree. The full text of the case paper included 105 paragraphs. After preprocessing the document, we got a vocab text including 1,290 non-repeated terms and a modeling file having 105 lines. A part of the vocab text and the modeling file is presented in Figure 6.

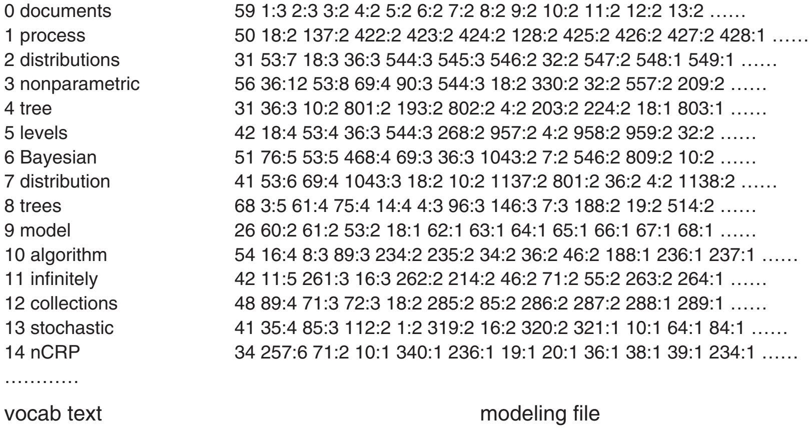
As shown in left part of Figure 6, each line of the vocab text starts with one term's serial number and then presents the corresponding term. So, the vocab text in our experiment has a total of 1,290 lines and ranks from 0. The modeling file has 105 lines, and as shown in right part of Figure 6, each line represents the number of terms, the serial number of every term and its frequency in one paragraph. Take the first line "59 1:3 2:3 3:2 4:2 5:2 6:2 7:2 8:2 […]" for an example: the number 59 means the first paragraph in the case paper contains 59 non-repeated terms, and formula 1:3 after 59 means the term whose serial number in the vocab text is 1 appeared three times in this paragraph. Similarly, formula 2:3 after 1:3 means the term "distributions" appeared three times in the first paragraph.

5.2 Parameter setting and experiment results

5.2.1 Parameter setting. As we mentioned earlier, the depth of the topic hierarchy tree and other parameters should be adjusted to meet the goal of analysis, so we had conducted experiments with different depths and parameters, then we found that the depth as three works the best. Meanwhile, we had set the nCRP parameter $\gamma = 1.0$ (sets the size of the inferred tree), the Griffiths-Engen-McCloskey (GEM) parameters were fixed at $m = 0.35$ (reflects the proportion of general words relative to specific words) and $\pi = 100$ (reflects how strictly we expect the documents to adhere to these proportions) with the number of sampling as 10,000.

After preprocessing the case paper, THC-DAT extracted the topic-words and hierarchy structure from the procedure results and got the topic hierarchy file for the case paper. The topic hierarchy tree has one root topic, five second-level topics and 11 third-level topics, so there are 11 topic paths in this tree. The 105 paragraphs are assigned to 11 topic paths, and each topic path has three layers of topic. Because each topic is a distribution of terms, THC-DAT selected top five terms representing each topic. In the processing, the vocabulary was restricted to the 1,647 non-repeated terms which is included in vocab text.

Then, a suitable threshold for merging the paragraphs which share a common topic should be selected based on the distribution of merging frequency. We found that



```
0 documents        59 1:3 2:3 3:2 4:2 5:2 6:2 7:2 8:2 9:2 10:2 11:2 12:2 13:2 ……
1 process          50 18:2 137:2 422:2 423:2 424:2 128:2 425:2 426:2 427:2 428:1 ……
2 distributions    31 53:7 18:3 36:3 544:3 545:3 546:2 32:2 547:2 548:1 549:1 ……
3 nonparametric    56 36:12 53:8 69:4 90:3 544:3 18:2 330:2 32:2 557:2 209:2 ……
4 tree             31 36:3 10:2 801:2 193:2 802:2 4:2 203:2 224:2 18:1 803:1 ……
5 levels           42 18:4 53:4 36:3 544:3 268:2 957:2 4:2 958:2 959:2 32:2 ……
6 Bayesian         51 76:5 53:5 468:4 69:3 36:3 1043:2 7:2 546:2 809:2 10:2 ……
7 distribution     41 53:6 69:4 1043:3 18:2 10:2 1137:2 801:2 36:2 4:2 1138:2 ……
8 trees            68 3:5 61:4 75:4 14:4 4:3 96:3 146:3 7:3 188:2 19:2 514:2 ……
9 model            26 60:2 61:2 53:2 18:1 62:1 63:1 64:1 65:1 66:1 67:1 68:1 ……
10 algorithm       54 16:4 8:3 89:3 234:2 235:2 34:2 36:2 46:2 188:1 236:1 237:1 ……
11 infinitely      42 11:5 261:3 16:3 262:2 214:2 46:2 71:2 55:2 263:2 264:1 ……
12 collections     48 89:4 71:3 72:3 18:2 285:2 85:2 286:2 287:2 288:1 289:1 ……
13 stochastic      41 35:4 85:3 112:2 1:2 319:2 16:2 320:2 321:1 10:1 64:1 84:1 ……
14 nCRP            34 257:6 71:2 10:1 340:1 236:1 19:1 20:1 36:1 38:1 39:1 234:1 ……
…………

vocab text                              modeling file
```

Figure 6.
Pre-processed result

$T^* = 1.0$ would make better merging effect, which indicates a smaller semantic deviation and better semantic fusion in the merging unit. Then we got the THC file for the case paper, whose logical structure is shown in Figure 7.

5.2.2 Analysis modes of our case. The main contribution of this study is that it provides a way for users to analyze document hierarchically. A few modes of how it could be employed for multi-grained document analysis are listed below:

(1) Navigation and browse mode: by skimming the topic hierarchy tree displayed in the navigating panel, one could have a clear comprehension about the case paper's topic structure, that it has three layers, five second-level topics, 11 leaf topics and 11 topic paths in total (as shown in Figure 8). So users may observe that the document is about five big topics or 11 small topics immediately.

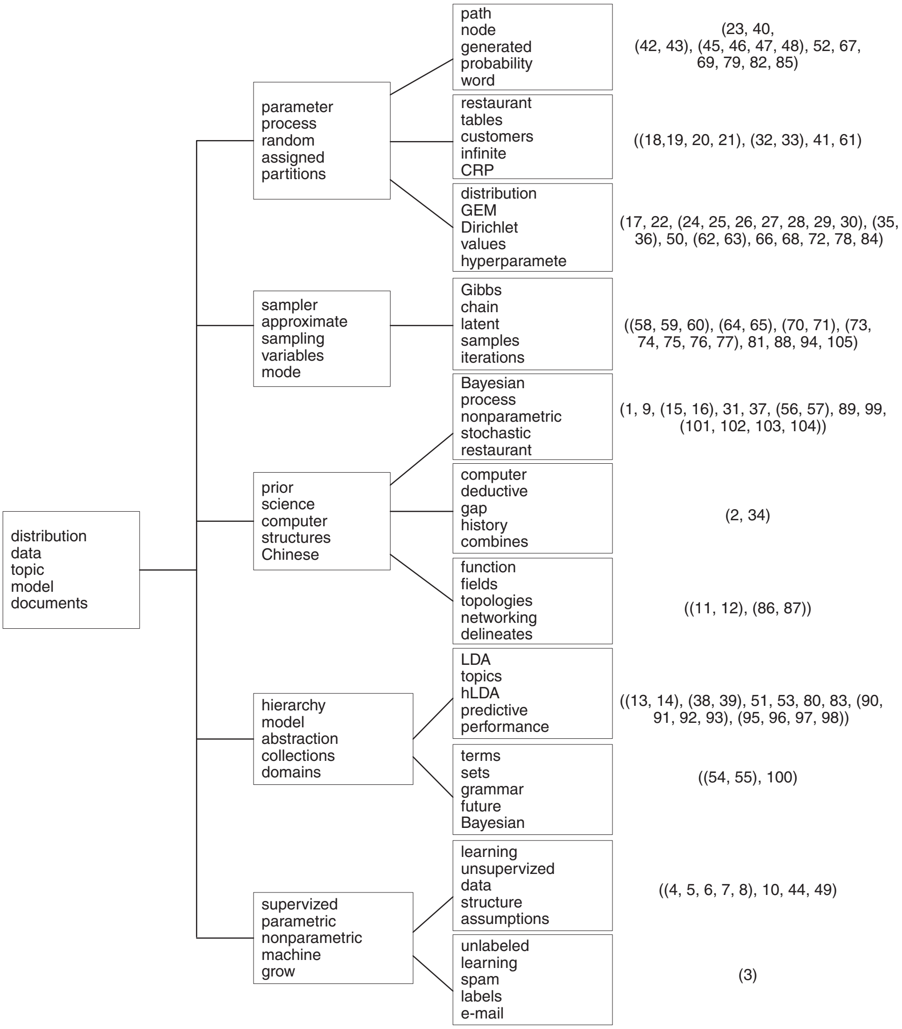
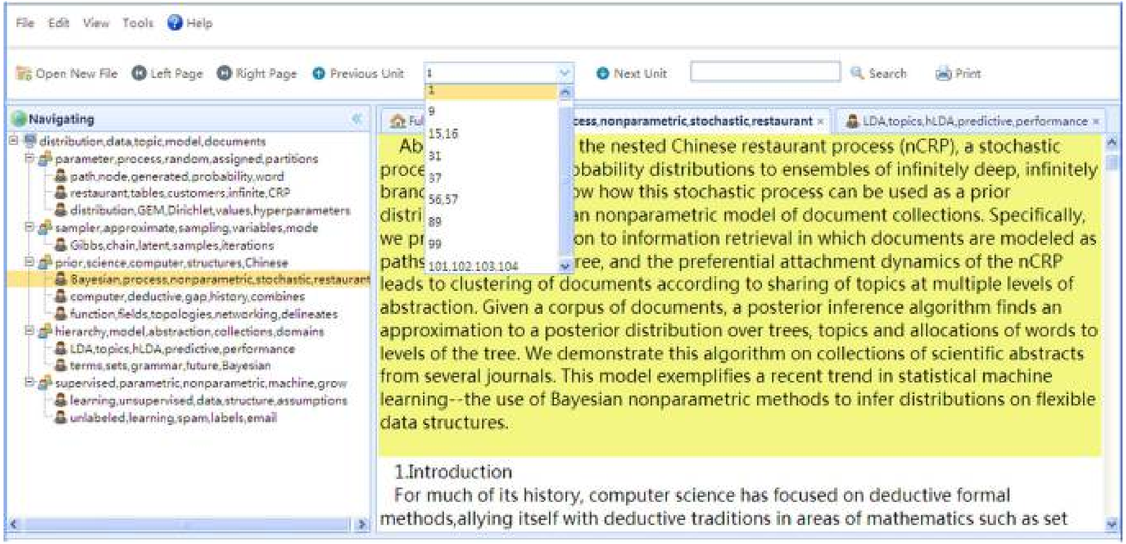

distribution data topic model documents
parameter process random assigned partitions
path node generated probability word
(23, 40, (42, 43), (45, 46, 47, 48), 52, 67, 69, 79, 82, 85)
restaurant tables customers infinite CRP
((18,19, 20, 21), (32, 33), 41, 61)
distribution GEM Dirichlet values hyperparamete
(17, 22, (24, 25, 26, 27, 28, 29, 30), (35, 36), 50, (62, 63), 66, 68, 72, 78, 84)
sampler approximate sampling variables mode
Gibbs chain latent samples iterations
((58, 59, 60), (64, 65), (70, 71), (73, 74, 75, 76, 77), 81, 88, 94, 105)
prior science computer structures Chinese
Bayesian process nonparametric stochastic restaurant
(1, 9, (15, 16), 31, 37, (56, 57), 89, 99, (101, 102, 103, 104))
computer deductive gap history combines
(2, 34)
function fields topologies networking delineates
((11, 12), (86, 87))
hierarchy model abstraction collections domains
LDA topics hLDA predictive performance
((13, 14), (38, 39), 51, 53, 80, 83, (90, 91, 92, 93), (95, 96, 97, 98))
terms sets grammar future Bayesian
((54, 55), 100)
supervized parametric nonparametric machine grow
learning unsupervized data structure assumptions
((4, 5, 6, 7, 8), 10, 44, 49)
unlabeled learning spam labels e-mail
(3)

Figure 7.
Logical structure of the THC file in this case


File Edit View Tools Help
Open New File
Left Page
Right Page
Previous Unit
1
Next Unit
Search
Print
Navigating
distribution,data,topic,model,documents
parameter,process,random,assigned,partitions
path,node,generated,probability,word
restaurant,tables,customers,infinite,CRP
distribution,GEM,Dirichlet,values,hyperparameters
sampler,approximate,sampling,variables,mode
Gibbs,chain,latent,samples,iterations
prior,science,computer,structures,Chinese
Bayesian,process,nonparametric,stochastic,restaurant
computer,deductive,gap,history,combines
function,fields,topologies,networking,delineates
hierarchy,model,abstraction,collections,domains
LDA,topics,hLDA,predictive,performance
terms,sets,grammar,future,Bayesian
supervised,parametric,nonparametric,machine,grow
learning,unsupervised,data,structure,assumptions
unlabeled,learning,spam,labels,email
Ful
cess,nonparametric,stochastic,restaurant ×
LDA,topics,hLDA,predictive,performance ×
1
9
15,16
31
37
56,57
89
99
101,102,103,104

Ab the nested Chinese restaurant process (nCRP), a stochastic
proce obability distributions to ensembles of infinitely deep, infinitely
branc ow how this stochastic process can be used as a prior
distri an nonparametric model of document collections. Specifically,
we pr on to information retrieval in which documents are modeled as
paths ree, and the preferential attachment dynamics of the nCRP
leads to clustering of documents according to sharing of topics at multiple levels of abstraction. Given a corpus of documents, a posterior inference algorithm finds an approximation to a posterior distribution over trees, topics and allocations of words to levels of the tree. We demonstrate this algorithm on collections of scientific abstracts from several journals. This model exemplifies a recent trend in statistical machine learning--the use of Bayesian nonparametric methods to infer distributions on flexible data structures.

1.Introduction

For much of its history, computer science has focused on deductive formal methods,allying itself with deductive traditions in areas of mathematics such as set

Figure 8.
Topic navigation and browse mode

Then, as shown in Figure 8, if any topic node in the topic hierarchy tree is clicked, the topic node and its corresponding text units will be highlighted, the browsing panel will be scrolled to the first unit automatically for browsing. Meanwhile, the unit dropping down list box also shows the sequence numbers of the paragraphs belonging to its corresponding text. So users can see the distribution of every topic just through the unit drop-down list box quickly. On being aware of the distribution of every topic, users will find out the main topic of the case paper in a moment. Specifically in our case, the first topic in the second level is the most important topic, because it includes 41 paragraphs that account for 40 percent of the full text. The last topic in the second level may be the least important topic because it only contains nine paragraphs. The second, third and fourth topic may have equal status because they include similar number of paragraphs. Further, by analyzing the topic terms and the distribution of the topics, users may have an idea that, the case paper is about stochastic process, Bayesian nonparametric statistics and topic model because there are professional terms in them, such as "topic model" in the root level, "process" in the first topic of the second level and "CRP," "GEM" and "Dirichlet" in its lower level topics. Similarly, "sampling" appears in the second topic of the second level while "Gibbs," as the specific sampling method of hLDA, in its lower level topic, "Bayesian" appears twice in the tree, one is in the fifth topic of third level and the other is in the ninth topic of third level, and "hierarchy" appears in the fourth topic of the second level while LDA and hLDA in its lower level topic.

As was mentioned above, for any topic, if user clicks the drop-down list box, the units being relative to the highlighted topic will be listed. As shown in Figure 8, the corresponding units of topic "Bayesian process nonparametric stochastic restaurant" are highlighted in the browsing panel and their location information or sequence number (1, 9, (15, 16), 31, 37, (56, 57), 89, 99, (101, 102, 103, 104)) are shown in the drop-down list box. From the topic terms and the units appearing in the methodology and discussion part of the paper, user may have the intuition that this topic is about the methods of the case paper.

Users can also identify the internal relations between units. Taking the units in the drop-down list box in Figure 8, for example, users can analyze these units from two aspects.

One is about the organization and basis of the whole paper. For example, the first unit of this topic is the abstract which introduces the contents, methods, application and significances of the case paper; then, the third unit includes the 15th and the 16th paragraph, in fact, the 15th paragraph is about the content arrangement of the case paper and is connected with the first unit, the 16th paragraph presents all the related theoretical basis of the case paper: stochastic process theory, Bayesian nonparametric statistics, CRP, stick-breaking process and Dirichlet process mixture; accordingly, the eighth unit (the 99th paragraph) summarizes the methods that this paper used. Through reading these units, users may quickly understand the methods used in the case paper and their relationships: nCRP is a new stochastic process which is developed by authors of the case paper and used as a prior distribution in a Bayesian nonparametric statistics, then, nCRP is used in the hierarchical topic model for analyzing document in terms of hierarchies of topics, after that, a posterior inference algorithm is used to define the underlying structure of topics, thus the topic hierarchies can be extracted finally.

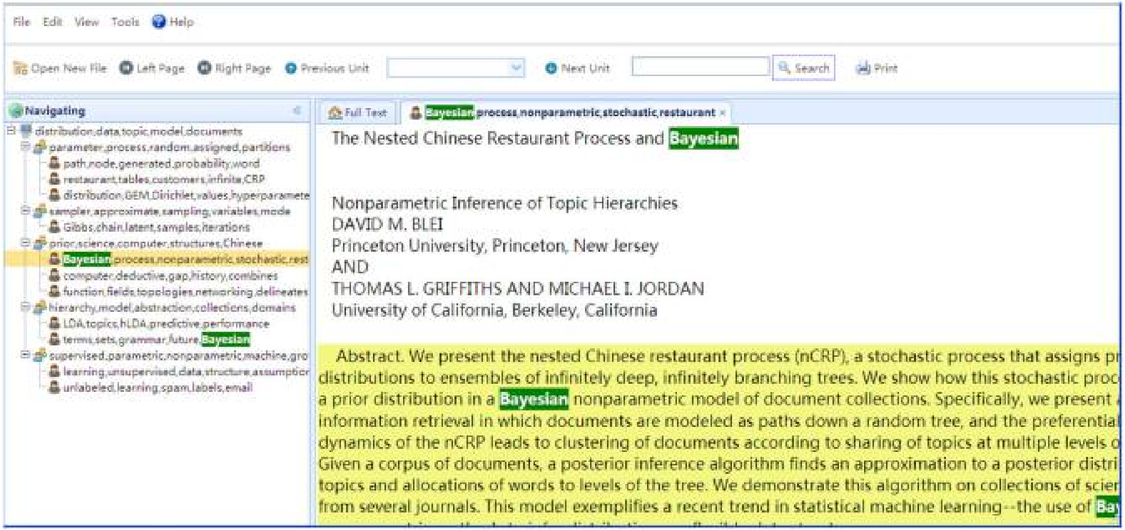
The other aspect is about the sequence of ideas of the case paper through which the author introduces the methods and principles step by step. For example, the second unit (the ninth paragraph) in the paper introduces the advantages of Bayesian nonparametric statistics; then the fourth unit (the 31st paragraph) takes a step further and illustrates the advantages and disadvantages of CRP often used in Bayesian nonparametric statistics, and prompts nCRP that can be used to create hierarchical topic model; furthermore, the fifth unit (the 37th paragraph) explain why nCRP can be used in probabilistic topic model, and the sixth unit (the 56th and the 57th paragraph) introduces the importance of posterior inference which is core problem of Bayesian nonparametric statistics and also hLDA; moreover, the seventh unit (the 89th paragraph) concludes that nCRP can yield different tree structures for different copra by the same posterior inference algorithm; and finally, the ninth unit including the 101st, the 102nd, the 103rd, the 104th paragraph, illustrates the advantages of every method and concludes the mutual improvements of these methods mentioned above. So it is very clear that every unit in this aspect is about one method used in the case paper, and they are related to each other.

Meanwhile, there are links connecting these two aspects, because the first one is for the whole and the second one is for the part. User can find out the close connection between these units just by shifting between them.

What is more, it is very interesting that most of these units are opening paragraph of sections. Specifically, the 16th paragraph of the third unit, the fourth unit, the fifth unit, the 56th paragraph of the sixth unit, the ninth unit are opening paragraph of section 2, section 3, section 4, section 5 and section 7, respectively. This also confirms our analysis above that this topic is about the methods used in the case paper and is the clue for understanding the relationships between methods and structure of the whole paper. However, from the distribution we can also find out it is just a clue rather than detailed discussion.

(2) Search mode: in this mode, one could input query term in the searching panel to search in both the topic hierarchy tree and the texts. For example, if a user is interested in "Bayesian nonparametric statistics" and inputs the term "Bayesian" into the searching panel, THC-DAT will search with the term and highlight the corresponding locations not only in the navigating panel but also in the browsing panel, as shown in Figure 9.

THC-DAT finds out total of 43 matches in the full text for "Bayesian," and also two matches in the topic hierarchy tree in navigating panel. So the user may realize immediately that the term "Bayesian" is very important for the case paper because it matches twice in the topic tree. The two topics including this term are "Bayesian, process, nonparametric, stochastic, restaurant" and "terms, sets, grammar, future, Bayesian," and they are both third-level topics. Obviously, from the position of "Bayesian" in these topics, the user may observe that the importance of the term "Bayesian" in the first relevant topic is much greater than that in the second relevant topic. Then, by scrolling the browsing panel, users may get more details about the distribution of query term in full text or just in corresponding units of one topic, and have insight of the query term. For example, the first relevant topic, as we discussed before, is about the methods the case paper used, in which Bayesian nonparametric statistics is the basic method, so the query term "Bayesian" appears in this topic and has got


File
Edit
View
Tools
Help
Open New File
Left Page
Right Page
Previous Unit
Next Unit
Search
Print
Navigating
distribution,data,topic,model,documents
parameter,process,random,assigned,partitions
path,node,generated,probability,word
restaurant,tables,customers,infinite,CRP
distribution,GEM,Dirichlet,values,hyperparamete
sampler,approximate,sampling,variables,mode
Gibbs,chain,latent,samples,iterations
prior,science,computer,structures,Chinese
Bayesian,process,nonparametric,stochastic,rest
computer,deductive,gap,history,combines
function,fields,topologies,networking,delineates
hierarchy,model,abstraction,collections,domains
LDA,topics,hLDA,predictive,performance
terms,sets,grammar,future,Bayesian
supervised,parametric,nonparametric,machine,gro
learning,unsupervised,data,structure,assumptior
unlabeled,learning,spam,labels,email
Full Text
Bayesian,process,nonparametric,stochastic,restaurant

The Nested Chinese Restaurant Process and Bayesian

Nonparametric Inference of Topic Hierarchies
DAVID M. BLEI
Princeton University, Princeton, New Jersey
AND
THOMAS L. GRIFFITHS AND MICHAEL I. JORDAN
University of California, Berkeley, California

Abstract. We present the nested Chinese restaurant process (nCRP), a stochastic process that assigns pr
distributions to ensembles of infinitely deep, infinitely branching trees. We show how this stochastic proc
a prior distribution in a Bayesian nonparametric model of document collections. Specifically, we present
information retrieval in which documents are modeled as paths down a random tree, and the preferential
dynamics of the nCRP leads to clustering of documents according to sharing of topics at multiple levels o
Given a corpus of documents, a posterior inference algorithm finds an approximation to a posterior distri
topics and allocations of words to levels of the tree. We demonstrate this algorithm on collections of scier
from several journals. This model exemplifies a recent trend in statistical machine learning--the use of Bay

Figure 9.
Search mode

the first rank. By clicking the second relevant topic, the user will have the idea that it is about the differences from other studies, and Bayesian nonparametric statistics is the sparkle, so the query term "Bayesian" also appears in this topic.

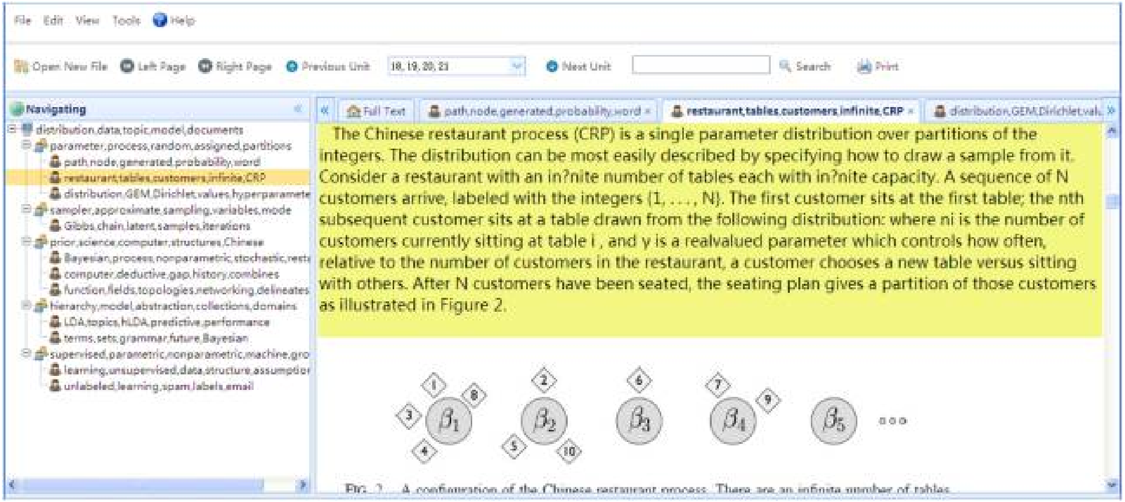
Besides the terms appearing in the topic hierarchy tree, THC-DAT also highlights all the locations where the term appears in the case paper, and they can also be combined with each other. In a word, THC-DAT provides multiple searching routes to help users have some deep insight about the query term in the document.

(3) Comparison mode: THC-DAT could help users comparatively study text through two patterns. One is, as mentioned above, comparing different semantic units under the same topic. When two units in the drop-down list box are clicked alternately, THC-DAT will automatically shift between their corresponding texts synchronously (as Figure 8 shown). So users could conveniently compare the units and find the relation and difference of intra-topic. The other is, comparing different topics and their corresponding texts so that users can find the relation and difference of inter-topic. When two or more different topics are clicked in turn, THC-DAT will automatically open new tab pages in turn to display the case paper and highlight the clicked topic's corresponding units, or switch to the corresponding tab page already opened. Figure 10 gives an example of comparing topics. For example, the second-level topic "parameter, process, random, assigned, partitions" has three child topics: "path, node, generated, probability, word," "restaurant, tables, customers, infinite, CRP" and "distribution, GEM, Dirichlet, values, hyperparamete." By observing and clicking these four topics, and browsing the corresponding semantic units, users may understand that, on one hand, the second-level topic, also the parent topic, is about the main model proposed in the case paper, the important method for the model, the basic theories and parameter setting for the model, and on the other hand, every child topic details one aspect of its parent topic. First, the topic "path, node, generated, probability, word" explains how to define the topic path and probability of a topic in hierarchical topic model. The terms "path," "node" and "probability" represent the content of this topic. To be clear, the hierarchical topic model is the most important model in the case paper and it is also the aim of the paper. Then, the topic "restaurant, tables, customers, infinite, CRP" introduces nCRP in detail, which is the main method used in hierarchical topic model for analyzing document in terms of hierarchies of topics. The terms "restaurant," "tables," "customers" and "CRP" represent the content of this topic. Further, the topic "distribution, GEM, Dirichlet, values, hyperparamete" details the basic theories and the parameter setting of hierarchical topic model. The terms "GEM," "Dirichlet" and "distribution," "hyperparamete" represent the content of this topic.

By comparing these topics and their corresponding texts, users will find that these topics are inseparably and closely related and form a clear logical chain. So users can get a better understanding about the structure, connotation, function and relationships of the topics, which makes users to read, learn and use the paper easier.

5.3 Discussion

The significant contribution of this study lies in that, it provides a new view of a full text document from its hierarchical topics and semantic text segments, and the tool THC-DAT enables more thorough and effective within-document analysis than



- distribution,data,topic,model,documents
  - parameter,process,random,assigned,partitions
    - path,node,generated,probability,word
    - restaurant,tables,customers,infinite,CRP
    - distribution,GEM,Dirichlet,values,hyperparamete
  - sampler,approximate,sampling,variables,mode
    - Gibbs,chain,latent,samples,iterations
  - prior,science,computer,structures,Chinese
    - Bayesian,process,nonparametric,stochastic,resta
    - computer,deductive,gap,history,combines
    - function,fields,topologies,networking,delineates
  - hierarchy,model,abstraction,collections,domains
    - LDA,topics,HLDA,predictive,performance
    - terms,sets,grammar,future,Bayesian
  - supervised,parametric,nonparametric,machine,gro
    - learning,unsupervised,data,structure,assumptio
    - unlabeled,learning,spam,labels,email

The Chinese restaurant process (CRP) is a single parameter distribution over partitions of the integers. The distribution can be most easily described by specifying how to draw a sample from it. Consider a restaurant with an in?nite number of tables each with in?nite capacity. A sequence of N customers arrive, labeled with the integers {1, . . . , N}. The first customer sits at the first table; the nth subsequent customer sits at a table drawn from the following distribution: where ni is the number of customers currently sitting at table i , and γ is a realvalued parameter which controls how often, relative to the number of customers in the restaurant, a customer chooses a new table versus sitting with others. After N customers have been seated, the seating plan gives a partition of those customers as illustrated in Figure 2.

FIG. 2. A configuration of the Chinese restaurant process. There are an infinite number of tables

Figure 10.
Comparison mode

existing tools. Based on this study, practical implications are discussed. It can help users deal with electronic documents from three aspects:

(1) First, it can help users to interactively read and analyze documents efficiently and effectively, in a topic hierarchy and semantic unit way. Within-document analysis is almost the first and most common task of users when using documents, while many documents such as academic papers have complex topic structure and big multi-professional vocabulary, but current within-DATs are relatively inefficient and ineffective for these characteristics. For example, DocuBurst (Collins *et al.*, 2009) visualizes documents based on term frequency and term hyponymy but not latent hierarchical topics, and Tiara (Liu *et al.*, 2012) does not consider the structure and content distribution of topics in a document. THC-DAT can support users to read and analyze documents by interactively navigating and searching the hierarchical topics, and browsing and comparing corresponding semantic text units, so it can enhance the efficiency and effect to analyze documents with topic hierarchy structure.

(2) Second, it can be used to support interactive, dynamical and exploratory learning about unfamiliar topics with little cognitive burden. New terms, topics or fields are one of the main obstacles in academic paper reading, especially in this interdisciplinary and rapid developing age. Some existing tools like DocuBurst (Collins *et al.*, 2009) do not consider new concepts, and some tools like WordTree (Wattenberg and Viégas, 2008) and TextArc (Paley, 2002) analyze new terms mainly based on term concurrence, which is too rough and inefficient to learn new things. THC-DAT can extract the latent topics and corresponding text segments in documents automatically, and help users discover and locate new topic or topic related to new terms quickly. Through navigating and browsing, users can get to know the corresponding text units and the function of the new things in the document, such as flaws of existing studies, key theory or technique to solve problem or explanation for the conclusions of the study. Furthermore, new topic can be learned quickly through comparing it to related topics in THC.

(3) Third, it helps to deepen user's understanding of a document. Existing tools emphasize almost just one aspect of document understanding. For example, DocuBurst (Collins *et al.*, 2009) can provide overall understanding, Xed (Hadjar *et al.*, 2004) focusses on layout but not semantic structure of the document, TextArc (Paley, 2002) visualizes the term distribution in the document, while ProfileSkim (Harper *et al.*, 2004) and overview scrollbar (Ko Mizoguchi, 2013) focusses on viewing local detail base on term distribution. THC-DAT can help users better understand documents by supporting overall understanding, structure analysis and viewing detail at the same time. The topic hierarchy and semantic unit organization gives the overall and structure view of document. Through navigating, browsing and comparing the hierarchical topics and the semantic units related to them, the overall and structure information can be observed, and through browsing and comparing semantic units based on their related topics, users can get to know the content and context of a certain unit.

Moreover, there are further implications of this research for text mining and data visualization, both of which are becoming an increasing necessity in academic libraries. The models in this paper can be utilized to build new knowledge mining, knowledge organization and knowledge visualization models of multi-topic documents, from hierarchical topics and their relevant texts aspect. For example, knowledge organization of e-books based on THC is one of our ongoing core researches till 2016.

6. Conclusion and future work

In this paper, we proposed an effective within-DAT THC-DAT for multi-topic documents based on THC information, making an attempt to apply hLDA and context information to analyze the inner topic of a document. In a journal paper analysis case, we conducted parameters to obtain topic hierarchy, and fixed context location threshold value to merge paragraphs in order to improve the analysis efficiency. Three typical analysis modes were also illustrated in the case. The case showed that THC-DAT can build a topic hierarchy tree for a document, show the corresponding semantic units for every topic in the topic hierarchy tree and search query term not only in topic hierarchy tree but also in the full text, so that users can search, browse and analyze documents effectively and efficiently, even without knowledge of the topics.

However, there are still many works to do in the future. hLDA provides a new idea for document analysis, but it needs users to specify the depth parameter and lacks autonomy. Efficient model to estimate and adjust the depth automatically in this tool needs to be further studied. Meanwhile, a user experience and user behavior study for this tool will be launched in September 2015. Considering the special style of and focus on medicine and biology papers, work on applicable version of this tool for papers in medicine and biology is also ongoing.

References

Agirre, E., de Lacalle, O.L. and Soroa, A. (2014), "Random walks for knowledge-based word sense disambiguation", *Computational Linguistics*, Vol. 40 No. 1, pp. 57-84.

Ahmed, S.Z., McKnight, C. and Oppenheim, C. (2006), "A user-centred design and evaluation of IR interfaces", *Journal of Librarianship and Information Science*, Vol. 38 No. 3, pp. 157-172.

Andrea, T. and George, K. (2013), "A segment-based approach to clustering multi-topic documents", *Knowledge and Information Systems*, Vol. 34 No. 3, pp. 563-595.

Bates, M.J. (2002), "The cascade of interactions in the digital library interface", *Information Processing & Management*, Vol. 38 No. 3, pp. 381-400.

Berg, S.A., Hoffmann, K. and Dawson, D. (2010), "Not on the same page: undergraduates' information retrieval in electronic and print books", *Journal of Academic Librarianship*, Vol. 36 No. 6, pp. 518-525.

Blei, D.M. (2009), "hLDA package", available at: www.cs.princeton.edu/~blei/topicmodeling.html (accessed March 25, 2014).

Blei, D.M., Griffiths, T.L. and Jordan, M.I. (2010), "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies", *Journal of the ACM*, Vol. 57 No. 2, pp. 1-30.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003a), "Latent Dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3 No. 5, pp. 993-1022.

Blei, D.M., Griffiths, T., Jordan, M.I. and Tenenbaum, J. (2003b), "Hierarchical topic models and the nested Chinese restaurant process", *Neural Information Processing Systems*, Vol. 16, available at: http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2003_AA03.pdf

Brosseau-Villeneuve, B., Nie, J. and Kando, N. (2014), "Latent word context model for information retrieval", *Information Retrieval*, Vol. 17 No. 1, pp. 21-51.

Brown, P.J. and Jones, G.J. (2001), "Context-aware retrieval: exploring a new environment for information retrieval and information filtering", *Personal and Ubiquitous Computing*, Vol. 5 No. 4, pp. 253-263.

Brown, G. and Yule, G. (1983), *Discourse Analysis*, Cambridge University Press, Cambridge.

Byrd, D. (1999), "A scrollbar-based visualization for document navigation", *Proceedings of the Fourth ACM Conference on Digital libraries, ACM*, pp. 122-129.

Collins, C., Carpendale, S. and Penn, G. (2009), "DocuBurst: visualizing document content using language structure", *Computer Graphics Forum*, Vol. 28 No. 3, pp. 1031-1039.

Crystal, D. (1991), *A Dictionary of Linguistics and Phonetics*, Wiley-Blackwell, Oxford.

Davis, P.M. and Price, J.S. (2006), "eJournal interface can influence usage statistics: implications for libraries, publishers, and project counter", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 9, pp. 1243-1248.

Dey, A.K. (2005), "Understanding and using context", *Personal Ubiquitous Computing*, Vol. 5 No. 1, pp. 4-7.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B. and Plaisant, C. (2007), "Discovering interesting usage patterns in text collections: integrating text mining with visualization", *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pp. 213-222.

Donoser, M., Wagner, S. and Bischof, H. (2010), "Context information from search engines for document recognition", *Pattern Recognition Letters*, Vol. 31 No. 8, pp. 750-754.

Du, L., Buntine, W., Jin, H. and Chen, C. (2012), "Sequential latent Dirichlet allocation", *Knowledge and Information Systems*, Vol. 31 No. 3, pp. 475-503.

Hadjar, K., Rigamonti, M., Lalanne, D. and Ingold, R. (2004), "Xed: a new tool for extracting Java application hidden structures from electronic documents", First International Workshop on Document Image Analysis for Libraries, IEEE, Palo Alto, CA, pp. 212-224.

Harper, D.J., Coulthard, S. and Yixing, S. (2002), "A language modelling approach to relevance profiling for document browsing", *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, ACM*, pp. 113-121.

Harper, D.J., Koychev, I., Sun, Y. and Pirie, I. (2004), "Within-document retrieval: a user-centred evaluation of relevance profiling", *Information Retrieval*, Vol. 7 No. 3, pp. 265-290.

Havre, S., Hetzler, E. and Whitney, P. (2002), "ThemeRiver: visualizing thematic changes in large document collections", *IEEE Transactions on Visualization & Computer Graphics*, Vol. 8 No. 1, pp. 9-20.

Hearst, M. (1995), "TileBars: visualization of term distribution information in full text information access", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, pp. 59-66.

Jiang, M.S., Liu, H.Y. and Li, X.Q. (1983), *The Basic Principle of College English Discourse Structure and Universal Law*, Weapons industry press, TianJin (in Chinese).

Ko, M., Daisuke, S. and Takeo, I. (2013), "Overview scrollbar: a scrollbar showing an entire document as an overview", *Human-Computer Interaction-INTERACT*, Springer Berlin Heidelberg, pp. 603-610.

Liu, S., Zhou, M.X., Pan, S., Song, Y., Qian, W. and Cai, W.L.X. (2012), "TIARA: interactive, topic-based visual text summarization and analysis", *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 3 No. 2, pp. 1-28.

Marchionini, G. and Komlodi, A. (1998), "Design of interfaces for information seeking", *Annual Review of Information Science and Technology*, Vol. 33 No. 3, pp. 89-130.

Messelodi, S. and Modena, C.M. (1996), "Context driven text segmentation and recognition", *Pattern Recognition Letters*, Vol. 17 No. 1, pp. 47-56.

Misra, H., Yvon, F., Cappé, O. and Jose, J. (2011), "Text segmentation: a topic modeling perspective", *Information Processing & Management*, Vol. 47 No. 4, pp. 528-544.

Moskovitch, R., Martins, S.B., Behiri, E., Weiss, A. and Shahar, Y. (2007), "A comparative evaluation of full-text, concept-based, and context-sensitive search", *Journal of the American Medical Informatics Association*, Vol. 14 No. 2, pp. 164-174.

Paley, W.B. (2002), "TextArc: showing word frequency and distribution in text", *IEEE Symposium on Information Visualization 2002 (InfoVis 2002)*.

Salton, G. and Allan, J. (1994), "Automatic analysis, theme generation, and summarization of machine-readable texts", *Science*, Vol. 264 No. 5164, pp. 14-21.

Stasko, J., Rg, C.G. and Liu, Z. (2008), "Jigsaw: supporting investigative analysis through interactive visualization", *Information Visualization*, Vol. 7 No. 2, pp. 118-132.

Sun, Y., Harper, D.J. and Watt, S.N.K. (2005), "Aiding comprehension in electronic books using contextual information", *European Conference on Research and Advanced Technology for Digital Libraries*, Springer-Verlag, pp. 504-506.

Suo, H., Nie, K. and Liu, Y. (2007), "Automatic summarization oriented topic partition", *Journal of Beijing University of Posts and Telecommunications*, Vol. 30 No. S1, pp. 14-17 (in Chinese).

Tanveer, J.S. and Tiwary, U.S. (2005), "Integrating relation and keyword matching in information retrieval", *Knowledge-Based Intelligent Information and Engineering Systems*, Springer Berlin Heidelberg, pp. 64-73.

van Dijk, T.A. (1983), "Discourse analysis: its development and application", *Journal of Communication*, Vol. 33 No. 2, pp. 20-43.

Venkatesh, R.K. (2013), "Legal documents clustering and summarization using hierarchical latent Dirichlet allocation", *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 2 No. 1, pp. 27-35.

Vo, D. and Ock, C. (2015), "Learning to classify short text from scientific documents using topic models with various types of knowledge", *Expert Systems with Application*, Vol. 42 No. 3, pp. 1684-1698.

Watt, R.J.C. (2015), "Concordance", (EB/OL), June 12, available at: www.concordancesoftware.co.uk/ (accessed June 12, 2013).

Wattenberg, M. and Viégas, F.B. (2008), "The word tree, an interactive visual concordance", *Visualization and Computer Graphics, IEEE Transactions on*, Vol. 14 No. 6, pp. 1221-1228.

Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F. and Singhal, A. (1999), "SCAN: designing and evaluating user interfaces to support retrieval from speech archives", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, pp. 26-33.

Zadrozny, W. and Jensen, K. (1991), "Semantic of paragraphs", *Journal Computational Linguistics*, Vol. 17 No. 2, pp. 171-209.

Zhang, Y., Gong, L. and Wang, Y. (2006), "Hierarchical subtopic segmentation of web document", *Wuhan University Journal of Natural Science*, Vol. 11 No. 1, pp. 47-50.

Further reading

Jiang, Y., Ding, X., Fu, Q. and Ren, Z. (2006), “Context driven Chinese string segmentation and recognition”, *Structural, Syntactic, and Statistical Pattern Recognition*, Springer Berlin Heidelberg, pp. 127-135.

Schwartz, M., Hash, C. and Liebrock, L.M. (2010), “Term distribution visualizations with focus +context”, *Multimedia Tools and Applications*, Vol. 50 No. 3, pp. 509-532.

Corresponding author
Quan Lu can be contacted at: mrluquan@hotmail.com

This article has been cited by:

1. Jing Chen Central China Normal University wuhan China Quan Lu Wuhan University Wuhan China Dan Wang Central China Normal University wuhan China Zeyuan Xu University of California Los Angeles United States MichaelS. Seadle Humboldt Universität zu Berlin Berlin Germany United States Laura Schilow Humboldt-Universität zu Berlin Berlin Germany . 2016. THC-DAT helps in reading a multi-topic document: results from a user-centered evaluation of a within-document analysis tool. *Library Hi Tech* **34**:4. . [Abstract] [PDF]