International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 2-No 6, 2012 – www.ijais.org

Legal Documents Clustering using Latent Dirichlet Allocation

Ravi kumar V
Dept. of IS&E
Research scholar, NIE
Mysore, India

K. Raghuveer
Dept. of IS&E
Faculty, NIE
Mysore, India

ABSTRACT

At present due to the availability of large amount of legal judgments in the digital form creates opportunities and challenges for both the legal community and for information technology researchers. This development needs assistance in organizing, analyzing, retrieving and presenting this content in a helpful and distributed manner. We propose an approach to cluster legal judgments based on the topics obtained from Latent Dirichlet Allocation (LDA) using similarity measure between topics and documents.

The developed topic based clustering model is capable of grouping the legal judgments into different clusters in effective manner. As per as our knowledge is concerned this is the first approach to cluster Indian legal judgments using LDA topic model.

General Terms

Documents Clustering, Similarity measure.

Keywords

Latent Dirichlet Allocation (LDA), Legal Judgments, Documents Clustering, cosine similarity.

1. INTRODUCTION

The outcome of WWW has made the courts around the world to provide online access to legal judgments of cases, for both past and present. Legal judgments are often complex in nature with multi-topical, containing carefully crafted, professional, domain-specific language and possess a broad and unevenly distributed coverage of legal issues. Clustering is an unsupervised data mining approach is extensively used in variety of situations. It automatically groups a collection into meaningful sub-groups; with a good document clustering method, computers can automatically organize a document corpus into a meaningful cluster hierarchy, which enables an efficient browsing and navigation of the corpus. An efficient document browsing and navigation is a valuable complement to the deficiencies of traditional IR technologies. Various clustering methods have also been proposed to the legal domain, but one of the interesting methods is to cluster the documents based on the relevant topic on which the documents are talking about. One of the most significant applications of topic segmentation is the Topic Detection and Tracking (TDT) task, as described in [1]. Much research has been carried out on topic segmentation. Many unsupervised, domain independent approaches [2, 3] exploit lexical cohesion information. The fact that related or similar words and phrases tend to be repeated in topically coherent segments and segment boundaries often correspond to a change in the vocabulary [4].Other approaches rely on complementary semantic knowledge extracted from dictionaries and thesauruses, or from collocations collected in large corpora, which use additional domain knowledge such as the use of hyponyms or synonyms [5,6,7,8].

In this paper our goal is to cluster the documents (legal judgments) into disjoint subsets of documents so that each subset represents documents which are most relevant to a topic obtained from Latent Dirichlet Allocation (LDA) for the given documents.

2. RELATED WORK

The ability to identify and partition a document into topics (segments) is important for many Natural Language Processing (NLP) tasks, including information retrieval, summarization, and text understanding.

Wei Xu et al. given an approach to cluster the documents based on the topics obtained from the corpus using the non-negative matrix factorization (NMF), in the latent semantic space where each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics [9].

Qiang Lu et al. describes a large scale soft clustering algorithm that relies on topic-segmentation for American legal documents using the Meta data of the legal documents [10].

Anna Huang, a wide variety of distance functions and similarity measures have been used for clustering text documents, such as squared Euclidean distance, cosine similarity, and relative entropy[11].

M. Saravanan et al. developed on ontology to retrieve legal judgments and to find summarization for some specific civil cases and it works based on the ontology constructed by human using legal corpus [12].

P. Berkhin, presented a complete survey on clustering algorithms [13].Clustering is an active area of research and a variety of algorithms have been developed in recent years. Clustering algorithm works based on the distance measure. Different distance measures give rise to different clusters. Thus, the distance measure is a significant means by which we can control the outcome of clustering.

3. OUR APPROACH

Document clustering is a process of organizing the documents into different clusters, such that the documents with in the cluster are more similar compare to the documents in the other cluster.

In this work our aim is to group the legal judgments into different cluster, so that it improves the information retrieval process by searching the required document within that cluster rather than searching the entire corpus.

The architecture of our approach to cluster Indian legal judgments into different clusters using topic documents similarity is shown in Fig. 1.0. The steps involved in this process are explained below.

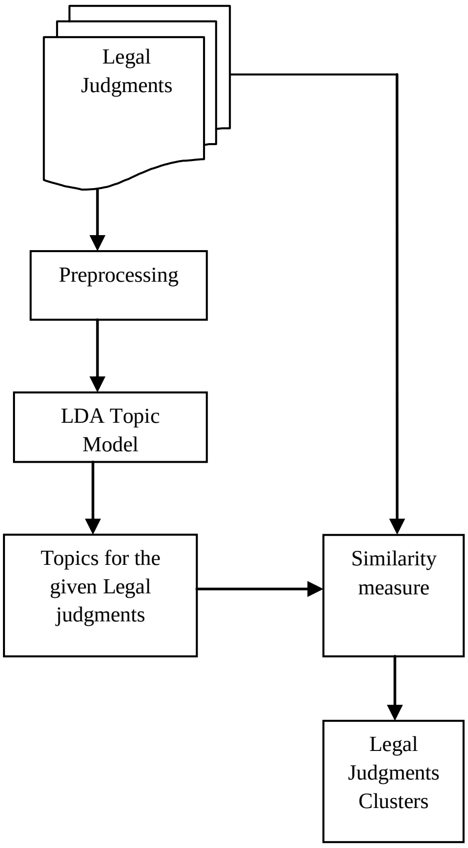

Fig. 1.0 Architecture to Cluster Legal Judgments

3.1 Legal judgments

Legal judgments are the text documents collected from [15], related to different civil cases. Legal judgments are very complex, in the sense that, there is no particular structure to express the opinion of the judges pertaining to different information about a case. The content of the legal judgment document can be divided into two main categories; they are

i) Summary about the case.

ii) The judgment part, gives the opinion given by the judge/s for the case in the form of free text.

In our work we have considered the judgment part of the document for clustering purpose.

3.2 Preprocessing

The judgment part of the legal judgment document is similar to other documents consists of stop words like is, of, an, etc. We remove these stop words to avoid in getting these stop words as topic terms. Similarly there are various legal terms they are common in all most all types of legal judgments documents and gives no information about the case, such terms are listed in consult with legal experts using legal judgments corpus. These terms considered as stop words and are removed from the input documents.

3.3 Legal Judgments into topics

Latent Dirichlet Allocation (LDA) is a probabilistic generative model used to model a collection of documents by topics, i.e., probability distributions over a vocabulary [14]. Given a vocabulary of W distinct words, a number of topics K, two smoothing parameters α and β, and a prior distribution (typically Poisson) over document lengths, this generative model creates random documents whose contents are a mixture of topics. With the use of LDA we break down the set of documents into topics. LDA uses a Dirichlet distribution to represent these topics and under the Dirichlet distribution these variables are independent of each other.

After the preprocessing we give all the documents to LDA and we get different topics based on this probabilistic model. Here we made an assumption that, the number of topics we get from LDA is equal the number of topics the corpus is describing.

3.4 Legal Judgments clustering using topics

Let D = $\{d_1...,d_N\}$ denote the set of documents to be cluster. K= $\{k_1,\ldots,k_M\}$ topics for the given corpus D documents obtained from the previous step.

For the clustering purpose, we use C= $\{ c_1,\ldots,c_M \}$ to denote the distinct cluster set that exists in the document collection D, of which c_k is one cluster consisting of documents representing a particular topic from different topics.

The common approach is to represent the documents to be clustered using vector-space model. A vector contains items from textual space, such as terms. The cosine similarity is applied to compute the similarity between two vectors *x*1 and *x*2 in the vector-space model, which is defined to be

$$\cos(x1, x2) = (x1 \cdot x2)/(\|x1\| \times\| x2\|),$$

Where $||x||$ is the length of a vector.

We have considered number of cluster is equal to the number of topics and hence we find the cosine similarity between each document with the topics and we place the document in to a cluster which is very close to a topic.

4. EXPERIMENTAL SETUP

4.1 Data set

We have considered Indian Legal Judgments from Kerala High court [15]. The data set consists of 120 documents from 6 different sub domains pertaining to civil case in India. Each document is labeled manually to which domain it belongs to.

4.2 Legal stop words list

When we started finding the topics using LDA for the given legal judgment documents, we observed that some of the legal terms which appears repeatedly in all most all types of cases and are less important from the experimental point of view, hence we thought of removing such words from the documents. With this we have taken legal judgment corpus and generated legal stop words list by taking the legal exports opinion.

4.3 Parameter for Gibbs sampling

Since we have considered 6 different types of civil case legal judgments in the corpus we have set K = 6 to match the number of anticipated clusters in the corpus. Following Blei et al. [14], we use α = 50/K and β = 0.1. Two additional parameters for the Gibbs sampling are the number of sampling and burn-in iterations, which we set to 30 and 200, respectively.

5. EXPERIMENT AND RESULT

We conducted the experiment to organize the legal judgments into different clusters using the cosine similarity between the legal judgments and topics obtained using LDA for the given corpus.

5.1 Legal Judgments-Topics Similarity

We compute the cosine similarity between each document in the corpus with topics and place the document into the cluster to which topic the document is closure. The table 1 shows cosine similarity measure between each document with the topics. The each row in the table shows document similarity over different topics of LDA and each column shows topic similarity over different documents. In the table, topic very close to the document has been made dark.

When we analyze the result, it shows that the documents from d_1 to d_{21} are grouped in to one cluster, as they are very close to one particular topic, topic-5 and when we cross checked about category of these documents manually we found that all are belongs to one particular case.

Table 1: Cosine similarity between documents to topics

documents	**1**	**2**	**3**	**4**	**5**	**6**
d1	0.022	0.002	0.011	0.004	**0.315**	0.012
d2	0.035	0.005	0.003	0.053	**0.154**	0.058
d3	0.027	0.007	0.010	0.002	**0.672**	0.016
d4	0.045	0.015	0.018	0.002	**0.120**	0.011
d5	0.032	0.014	0.013	0.012	**0.210**	0.003
d6	0.105	0.005	0.001	0.006	**0.484**	0.002
d7	0.021	0.001	0.005	0.003	**0.042**	0.014
d8	0.048	0.008	0.002	0.004	**0.068**	0.012
d9	0.011	0.001	0.027	0.069	**0.199**	0.002
d10	0.013	0.003	0.004	0.008	**0.065**	0.006
d11	0.063	0.003	0.007	0.002	**0.321**	0.005
d12	0.052	0.002	0.003	0.001	**0.493**	0.080
d13	0.042	0.003	0.014	0.026	**0.086**	0.015
d14	0.018	0.002	0.001	0.087	**0.089**	0.07
d15	0.013	0.003	0.018	0.006	**0.285**	0.012
d16	0.015	0.011	0.012	0.009	**0.339**	0.038
d17	0.073	0.023	0.017	0.007	**0.349**	0.029
d18	0.033	0.003	0.020	0.015	**0.056**	0.008
d19	0.057	0.017	0.025	0.066	**0.129**	0.045
d20	0.043	0.003	0.004	0.009	**0.087**	0.021
d21	0.025	0.020	0.017	0.019	**0.206**	0.038

The top 10 terms of each topic obtained from LDA topic model is shown in Fig. 2.0

Topic 1	Topic2	Topic3
tax	days	assessment
assessee	enquiry	income
sales	period	tax
authority	leave	officer
rate	service	assessing
turnover	rules	business
goods	provisions	deduction
sold	procedure	return
exemption	permit	amount

Topic4	Topic5	Topic6
compensation	building	accused
insurance	rent	cheque
vehicle	tenant	complainant
motor	landlord	negotiable
company	control	instruments
accident	premises	execution
driver	schedule	offence
bus	passed	trial
autorickshaw	authority	prosecution

Fig 2.Top 10 topic terms from the LDA topic model for the given Legal Judgments.

5.2 Evaluation Metrics

The performance of our approach to cluster Legal Judgments documents has been evaluated using two metrics i) entropy and ii) purity, based on the true class memberships in a document set.

i) Entropy: It is a function of the distribution of classes within the resulting clusters,

Given a particular cluster, C_r, of size n_r, the entropy of this cluster is defined as

$$E(\mathrm{C_r}) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}log\frac{n_r^i}{n_r} \quad (1)$$

Where q is the number of classes in the data set, and n_r^i is the number of documents of the i^{th} class that were assigned to the r^{th} cluster.

The entropy of the entire clustering solution is then defined as the sum of the individual cluster entropies weighted according to the cluster size.

$$\mathrm{Entropy} = \sum_{r=1}^{k}\frac{n_r}{n}\mathrm{E(C_r)} \qquad (2)$$

An ideal clustering solution will result in clusters that include documents from only a single class, in which case the entropy will be zero. In general, the smaller the entropy values, the better the clustering solution is.

ii) Purity: It is a function of the relative size of the largest class in the resulting clusters.
The purity of a cluster is defined as

$$P(C_r) = \frac{1}{n_r} \max_i \left(n_r^i\right) \qquad (3)$$

It is the number of documents of the largest class in a cluster divided by the cluster size. The overall purity of the clustering solution is obtained by taking a weighted sum of the individual cluster purities and is given by

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(C_r) \qquad (4)$$

In general, the larger the value of purity, the better the clustering solution is.

5.3 Results from Legal Judgments Collection

The results for organizing legal judgments into different cluster using our approach is shown in table 2.

Table 2: Performance Metrics for Clustering Legal Judgments

Cluster No.	No. of Classes	Purity	Entropy
C1	1	1	0
C2	1	1	0
C3	1	1	0
C4	2	0.529	0.214
C5	2	0.826	0.099
C6	1	1	0

The result shows that the overall purity of the clustering solution is greater than the entropy of the entire clustering solution and hence qualities of the resulting clusters are remarkable.

6. CONCLUSION AND SCOPE FOR FUTURE

An attempt has been made to organize Legal Judgments in to different cluster using cosine similarity between Legal judgments and topics, to improve the IR performance. The result shows that the clustering of Legal Judgments can be achieved by just finding the cosine similarity between Legal Judgments and topics. In future we are planning to segment the given Legal Judgment to generate summary of the legal judgment.

7. ACKNOWLEDGMENT

We would like to thank Mr. Shivakant Aradya, advocate, for the assistance and guidance given to us in preparing legal stop words list and for analyzing topic terms generated from LDA.

8. REFERENCES

[1] J. Allen, et al. “Topic detection and tracking pilot study final report”. In Proc. of the DARPA Broadcast News Transcription and understanding Workshop, 1998.

[2] Marti Hearst. “Texttiling: Segmenting text into multi-paragraph subtopic passages”. Computational Linguistics, 1997, Vol. 23.Pages 33–64.

[3] M. Utiyama and H. Isahara. “A statistical model for domain-independent text segmentation”. In Proc. of the ACL 2001, pages 499–506.

[4] M. Shafiei and E. Milios. “A statistical model for topic segmentation and clustering”. In Proc. of Canadian AI'08.

[5] D. Beeferman, A. Berger, and J. Lafferty. “A model of lexical attraction and repulsion”. In Proc. of the ACL, pages 1997, pages 373–380.

[6] F. Choi, P. Wiemer-Hastings, and J. Moore. “Latent semantic analysis for text segmentation”. In Proc. of EMNLP, 2001, pages 109–117.

[7] H. Kozima. Text segmentation based on similarity between words full text. In Proc. of the ACL, pages 286–288, 1993.

[8] H. Kozima and T. Furugori. “Similarity between words computed by spreading activation on an English dictionary”. In Proceedings of the ACL, 1993, pages 232–239.

[9] Wei Xu, Xin Liu and Yihong Gong. “Document Clustering Based On Non-negative Matrix Factorization”. In Proc. of *SIGIR’03* July 28–August 1, 2003, Toronto, Canada.Pages267-273

[10] Qiang Lu, William Keenan, Jack G. Conrad and Khalid Al-Kofahi. “Legal Document Clustering with Built-in Topic Segmentation”. In Proc. of CIKM’11, October 24–28, 2011, Glasgow, Scotland, UK. Pages 383-392

[11] Anna Huang. “Similarity Measures for Text Document”. In Proc. of *NZCSRSC 2008*, April 2008, Christchurch, New Zealand.

[12] M. Saravanan., B. Ravindran and S. Raman. “Using Legal Ontology for Query Enhancement in Generating a Document Summary”. In Proc. of JURIX 2007, 20th International Annual Conference on Legal Knowledge and Information Systems, Leiden, Netherlands, 13-15th Dec 2007.Pages 171-172.

[13] P. Berkhin. “A survey of clustering data mining techniques”. Grouping Multidimensional Data 2006, pages 25–71.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent Dirichlet allocation”. Journal of Machine Learning Research Vol.3 (2003) 993-1022.

[15] http://www.keralawyer.com/asp/sub.asp?pageVal=judgements