www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 3 March 2017, Page No. 20634-20637
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i3.43

Restricted Legal Dictionary generation for Legal Domain Specific Mining

B.V. Rama Krishna[1], B. Basaveswar Rao[2], K.Gangadhar Rao[3],K.Chandan[4]
[1]Research Scholar, Dept. of CSE
[2]Professor Computer Center
[3]Professor CSE Department
[4]Professor Dept. of Statistics
Acharya Nagarjuna University

ABSTRACT
The Indian legal domain maintains largest heterogeneous multi domain based text corpora. In order to document specific search domain specific dictionaries are widely acceptable compared to whole legal dictionary. Many domain specific dictionary extractions proposed by various authors for business and scientific domains. In this paper we proposed a two stage domain specific dictionary for legal domain. The approach helps to construct Restricted Legal Dictionary which is domain specific in nature. The dictionary constructed supports text mining mechanisms over specific crime (dowry) based legal documents and improves the BoW construction with higher domain specific terminology. The Restricted Legal Dictionary constructed in this paper improves the efficiency of similarity measures during document comparison techniques
.

Key Words: - Restricted Legal Dictionary (RLD), Domain Specific Dictionary, Domain ontology, Human Expert

1. INTRODUCTION

The increased demand on knowledge extraction from large text corpora through web portals placed many challenges to search engines. To search specific topic based information from documents search engines rely on dictionaries, ontologies and linguistic tools. Mining domain specific texts using glossaries to identify concepts and terms supports automatic Ontology generation [3]. The grouping of new domains purely based on concepts identified. However domain expert review is an essential phase for these mechanisms.

Different provenances maintain diverged set of vocabularies through dictionaries [2]. To extract information by provenance based dictionaries utilize statistical models with rule based refinement process [4]. Domain Ontology construction from available domain specific dictionaries is also widely used to generate *Restricted Dictionaries* [5]. The Restricted Dictionary (RD) is a limited set of lexicon with high degree of domain specific word list. RDs are compact in nature but effective in text mining applications over specific domains. Sometimes they are referred as domain specific dictionaries specifically treated as subset of dictionaries [1][6] which restricts the scope of glossaries within rigid boundaries fixed according to criteria of domain. The RDs are flexible compared to language dictionaries and they can be constructed on spatial or multimedia data also [7]. The domain specific dictionaries constructed over such data organized into nested hierarchical structures. The strong theme behind the process of construction of RLDs is to reduce unnecessary glossary verification and improve search reliability. The availability of third party tools like CRYSTAL [8] in market satisfies the needs of corporate and scientific sectors in generating domain specific dictionaries. But still there is a need to identify new mechanisms which reduce complexity and improve flexibility to generate Restricted Dictionaries for legal domain. The opinion based word ontologies creation used in social media based dictionaries creation [9].

In section 2 methodology of our work described. In section 3 the two staged process of RLD construction explained. Section 4 depicts results and analysis followed by conclusion and future work.

2. METHODOLOGY

2.1 Data Extraction

The Indian Legal Dictionary downloaded from web site (http://www.SupremecourtIndia.com/ library/LD.txt) for this work. Legal dictionary maintains a large set of glossaries related to all IPC Sections and Acts. This dictionary acts as base for construction of RLD related to dowry domain.

2.2 Human Expert Evaluation

DOI: 10.18535/ijecs/v6i3.43

The legal dictionary subjected to domain expert verification. During this evaluation process legal expert identifies domain specific (Dowry) legal words from the legal dictionary. The advantage of human expert evaluation improves the provenance based legal vocabulary rate in restricted legal dictionary.

3. RLD CONSTRUCTION

The process of RLD for dowry domain constructed in a two phased manner as shown
in Figure 1. During phase-1 legal domain expert identifies the legal words related to dowry specific crimes among the legal dictionary vocabulary belong to Indian territory.

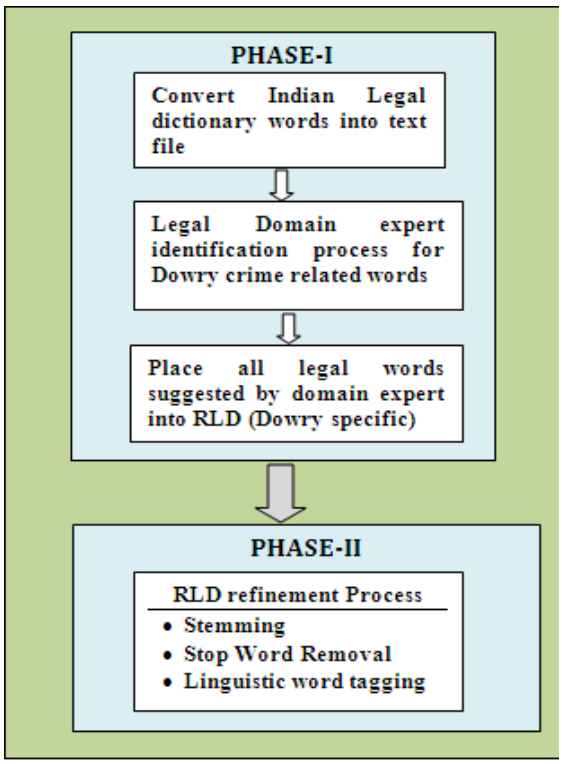

Figure 1. RLD construction Process

The process of human evaluation is performed by senior lawyers who are well knowledge as well as having higher professional experience in dowry domain. Once the Restricted Legal Dictionary (Dowry) with domain expert suggested words constructed voting begins. The selection of domain (dowry) specific legal words is based on voting mechanism with positive and negative feedback validation procedure. All the domain specific legal words suggested by domain expert forwarded to lawyers community for feedback purpose. The words with highest positive feedback considered as most widely used domain specific words, above average positive feedback indicates that those words are occasionally used by judicial proceedings and are having significant meaning. The words with high negative feedback are not used in judicial territories hence treated as insignificant.

Once all the significant words with positive feedback identified a word list constructed which is RLD for dowry related crimes. This word list passed to phase-2 for refinement process. In phase-2 *stemming* applied over RLD to conflation of words. *Stop-word removal* and *linguistic word tagging* applied sequentially to finalize legal words. This improves and strengthens the RLD vocabulary as well as concept density of dictionary.

4. RESULTS & ANALYSIS

There are 245 legal words suggested by domain expert for dowry specific crimes as entries for RLD. These legal words are validated by lawyer's community through feedback mechanism based on their usage in Indian judicial territories.

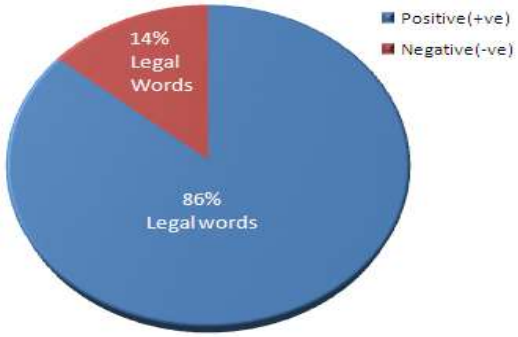

Figure 2. Provenience based usage Feedback on RLD

The above figure shows the percentage of positive feedback and negative feedback for RLD vocabulary. Around 34 legal words are rarely or not used in legal proceedings in local territories. Hence 214 legal words are considered as most significant legal words specifically related to dowry domain in legal domain. These legal words considered as RLD and used for dowry related judicial document mining. The least significant score of negative feedback indicates that the RLD constructed using this methodology is highly reliable and support domain expert scale knowledge for Information Retrieval mechanisms.

4.1 Performance Analysis

The Restricted Legal Dictionaries are compact and consumes less memory space. They are highly topic sensitive hence effectively applied for Bag-of-Words construction, document similarity score estimations and Information retrieval applications. In this paper general Legal Dictionary contrasted with RLD over time complexity measures to handle 200 dowry crime

related documents during Information Retrieval mechanisms. As shown in Table 1 RLD shows improved performance over general Legal Dictionary in handling domain specific text documents.

Table 1. Performance Metrics

Dictionary	Memory Usage	Process Time	Domain Knowledge	Updating Cost
Legal Dictionary	Large(3.66 MB)	High (45 Sec)	Heterogeneous	High
RLD	Small (926 KB)	Low (20 Sec)	Homogenous	Low

4.2 RLD for Dowry domain

The typical RLD constructed from this work is given below

abet	ablaze	abus	abuse
accompani	accus	act	action
admission	advis	advocate	aid
alimony	anti	anti-dowri	appeal
admit	article	assault	assist
attack	attempt	bear	beat
behavior	bench	bride	brief
bribe	bring	broad	burn
case	coercion	commit	common
consid	constitut	consum	continu
concubine	contributor	crime	criminal
cruelti	cumul	daughter	death
decree	defense	defend	demand
depress	desper	disguising	dissatisfact
domest	dous	dowri	econom
emot	encourage	extend	extract
famili	fire	forc	form
girl	grounds	hang	harassment
harm	helpless	high court	higher
hostag	husband	impact	includ
instanc	intention	intimid	issue
judgment	jurisdiction	kerosen	law
lead	leave	left	legal
liability	main	major	marri
marriage	matter	meet	murder
newlywed	notice	object	observe
occur	offence	order	party
person	petition	physic	poison
predomin	prior	prohibition	properti
protect	rang	rate	reality
reason	recogn	reduce	registry
relative	respondent	revie	right
secure	session	sexual	show
signific	situation	social	specific
specimen	state	statute	stayorder
subjudice	suicide	supreme	target
threat	tortur	type	typical
tolerate	unable	valuable	verbal
victim	violence	volit	vulner
wide	wife	womb	woman
women	yield	year	young

Figure 3. RLD vocabulary

Hence RLD stored in text format as external file can be easily adoptable to Data Mining tools and Java modules during text mining applications. The web interfaces can also be easily tailored with RLD for OLAP procedures.

CONCLUSION

The two-staged process presented in this paper is highly recommended for quick and less complex oriented RLD generation for a specific domain. The RLD maintains good domain information gain with reliable provenance based vocabulary. The feedback process conducted over domain expert suggested RLD identifies the micro outliers exist in domain specific dictionaries. Finally a double filtered approach applied in this work improved the legal words relativeness to specific domain. In future machine learning techniques applied to improve the quality of proposed framework and automate the process of knowledge learning mechanisms to support RLD generation with less human intervention.

REFERENCES

[1] Myeong So Kim et. al. "Performance Evaluation of Domain Specific sentiment Dictionary Construction Methods for Opinion Mining", IJDTA, Vol. 9,pp:257-268, 2016.

[2] P. Hall et. al. "When Words Matter Most: Tailoring Domain Specific Dictionaries With Decision Analytics", IST-Conference, 2015.

[3] V. Parekh et. al. "Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies", Maryland University, Baltimore, 2014.

[4] David Fisher et. al. "CRYSTAL Inducing a Conceptual Dictionary",University Press, Massachusetts, 2012.

[5] Tabraiz Anwer et. al. "Automatic Generation of Domain Specific Keywords", SZABIST Conference, JISRC, Vol. 10, Karachi, July, 2012.

[6] Y. Hayase et. al "building Domain Specific Dictionaries of Verb-Object Relation from Source Code", KAKENHI-Project,Osaka University Grants,2010.

[7] P. Agathangelou et. al. "Mining Domain Specific Dictionaries of Opinion Words",Hellenic University Press,2009.

[8] Patrick Drouin et. al. "Detection of Domain Specific Terminology Using Corpora Comparison", 2002.

[9] Ellen Riloff et. al. "Automatically Constructing a Dictionary for Information Extraction Tasks",National Conference on AI, MIT Press, 1993.

AUTHORS PROFILE

Currently doing research in data mining from Acharya Nagarjuna University. His areas of interest are Data Mining, E-Commerce, and Computer Graphics. He has 14 years of teaching experience and a lifetime member of CSI.

He is currently working as Director of University Computer Center (ANRU). His areas of interest are Computer Networks, Cloud Computing, Data Mining and Information Security. He is the author of many technical papers and guiding university scholars in CSE department.
Teaching Experience : 35 years
Research Experience : 10 years

Author is currently acting as Chairman of University (ANRU) CSE department. His areas of interest are computer networks, Information Security, Cloud Computing and Data Mining. He is the author of many technical papers and guiding university scholars in CSE Department.
Teaching experience : 30 years
Research experience : 10 years

Author is currently acting as Head of the statistics department of ANRU. His areas of interest are Statistics, Data Mining, Turing Machines and Automata Theory. He is author of many technical papers and guiding university scholars from statistics department.
Teaching Experience : 38 years
Research Experience : 15 years