IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. IV (Jul.-Aug. 2016), PP 58-63
www.iosrjournals.org

Enhancement of Bag-of-Words for Legal documents using Legal Statute

B.BasaveswarRao[1], B.V.RamaKrishna[2], K.GangadharRao[3] and K.Chandan[4]
[1] *Computer Centre, Acharya Nagarjuna University Guntur, 522501, A.P, India*
[2] *Dept. of Computer Science & Engineering, Acharya Nagarjuna University Guntur, 522501, A.P, India*
[3] *Dept. of Computer Science & Engineering, Acharya Nagarjuna University Guntur, 522501, A.P, India*
[4] *Dept. of Statistics, Acharya Nagarjuna University Guntur, 522501, A.P, India*

Abstract: *In this paper Legal statute related to dowry acts has been processed to obtain a distinct set of legal keywords which don't have a common occurrence in day to day dowry case judgments. This effort coupled with the knowledge of legal expert would give a very much broadened scope for the BoW. These keywords are very rich in concept and well connected to the domain of dowry acts. The earlier work [22] constructed BoW for dowry case notes of judgments. Current work tries to improve the BoW by widening the scope of dowry related cases. This enriches the Bag-of-Words with high probability legal terms taking precedence over low probability non-legal terms. The enriched BoW set when put through any of the similarity measures or machine learning techniques is bound to give better results when compared to earlier BoW[22].*
Keywords: *Bag-of-Words, Legal Statue, Dowry-Act, Enhancement.*

I. Introduction

In the contemporary times where both transactional and analytical data has become so voluminous and unmanageable because of the penetration of the Internet. Likewise the amount of legal information is also voluminous and is available online. To access the legal documents is becoming more and more difficult in the context of synthesized semantic knowledge [8]. The best solution for the representation of documents is BoW for applying classification and clustering techniques. There is always a need for enrichment of BoW [17, 19] for better results. Most Information Retrieval systems match documents syntactically. Computing Semantic Textual similarity [2] uses word ontologies, thesaurus and topic keyword mapping, probabilistic topic model techniques, natural linguistics [6, 12, 17]. The user query may not contains all the relevant document terms, more topic terms reside on corpus or external knowledge sources [14]. The Bag-of-Words approach is a common technique for estimating similarity among documents and query text. The Wikipedia, Word Net and Thesaurus [7, 10, 22] are used for Bag-of-Words construction with high topic enriched keywords. These external knowledge sources [13, 16, 18, 20] have been used by several authors to enrich Bag-of-Words so that the user queries expansion as well as document expansion [19, 21] can be handled more efficiently. They yield high topic sensitive ranking. A phased pre-processing such as Stemming, Stop-Word removal and Normalization improve document representation but suffers from drawbacks like order of terms maintenance failure, treating synonyms as different words and ambiguity [9] among grouped words as single units.

In this paper the two primary sources for constructing enriched BoW have been identified as Legal statute pertaining to dowry acts [304B, 498, 256] and the vast knowledge accrued by the legal experts over a period of time. The reason for choosing Legal Statute as the one of the external knowledge source [4, 5] for constructing the enriched BoW is that statute happens to be the basis for the different sections of IPC (Indian Penal Code). The enriched BoW thus created is a semantic BoW which can be used as a major source of metadata for the researchers whose research area happens to be dowry cases.

The Legal documents are represented as BoW representation for operation Machine Learning techniques like Classification, Clustering and finding Similarities between the documents [1]. All these studies are based on quality of BoW representation. So it is need to perform enrichment operations on BoW vector representation [23]. The exclusivity quantification of the BoW is estimated with Word Similarity metric. Word similarity is to find the similarity between the four BoW's (SBoW, LBoW, HBoW, and EBoW) and to that of legal dictionary. Many researchers estimated Word Similarity among BoW and documents using Cosine Similarity statistical measurement. The percentage score represents how much concentration of terms related to specific theme exists. The objective of the paper is twofold the first part handles enrichment of BoW and the second part focuses on exclusivity quantification by using cosine similarity measure.

The rest of this paper is organized as follows. In section 2 various authors contribution related to BoW enrichment techniques are discussed. Section 3 handles data description and methodology for BoW enhancement is presented in Section 4. Section 5 is dedicated for comparative analysis and results. Finally conclusion and future scope discussed in Section 6.

II. Literature Study

Zichao et.al proposed a cluster based representation enrichment method to overcome the problems related to shortness and sparseness of short text data. A topic relevant vector adopted with TF-IDF representation maintains topic relevance's derived from knowledge embedded in short texts. Finally using hierarchical clustering algorithm with purity control topics clustered. The results show improvement over original representation. With integration of external knowledge resources CREST yield high quality of clusters.

Erich Schweighhefer et. al introduced lexical ontologies with user feedback to improve Boolean search with query expansion. The feasibility of this approach is based on Lexical Ontology and Knowledge base.

A.Ranjanpal et.al proposed a technique to find meaning of words using word sense disambiguation using supervised and unsupervised learning. The work focused on limitations of supervised approach by using "modified LESK" and "Bag-of-Words". New bag of words constructed using learning methods.

Muyu Hang et.al proposed a triple [argument, predicate, argument] knowledge based document enrichment framework. This model performed more effectively than conventional enrichment techniques like external knowledge base and Ontologies. Stiffan Mizzaro et.al proposed method to enrich short texts based on semantic values extracted from web pages of same context. External Wikipedia knowledge base used to categorize topics. A predefined set of Wikipedia categories used in this model. This approach also used to identify new topics in text corpora's. Peng wang et. al proposed bit term model which constructs topic-keyword graph with link analysis. Once keywords are identified all these are appended to short text. BTM improve tf-idf score, also a new weighted Sim-Rank proposed to measure affinity among nodes.

V.Rekha et. al proposed pseudo relevance feedback technique which reform query based on central theme of the document. The expansion terms obtained from Eqi-Frequency partitioning using TF-IDF scores. Also rule based statistical models selects group of words. Claudio Carpineto and Giovanni Romano performed a survey on automatic query expansion and overviewed on interactive query refinement, relative feedback, concept lattice based, latent semantic indexing. Document Centric and Theme Centric approaches are experimentally analyzed by them to improve Automatic Query Expansion.

Hyun Dukkin et. al implemented probabilistic topic model which uses frequent word patterns to capture semantic associations between words. This model improves classification task. They applied on PLSA and LDA topic models. Their work shows improvement in topic modeling performance.

III. Data Description

The legal statute of dowry act obtained from http://www.justiceindia.com/volumes/statutes/sec/498 [25]. The obtained document is in PDF format. This document is converted to text document for further pre-process using previous work to generate Legal Statute BoW (LBoW). The Standard BoW (SBoW) generated from previous work [24]. The Bag-of-Words suggested by the Legal expert after thorough examination of SBoW and LBoW, considered as HBoW for this experiment of construction of enriched BoW. 'Legal Dictionary' available on the web http://www.supremecourt india.com/Library/Legal Vocabulary/Legal Dictionary.pdf [25] downloaded for performing word similarity comparison experiment of these BoWs. This Legal Dictionary is a collection of Indian Legal terminology. From this dowry related words are extracted into a text file. This text file acts as Legal Dictionary for conducting Similarity measure with set of Bow's.

IV. Methodology

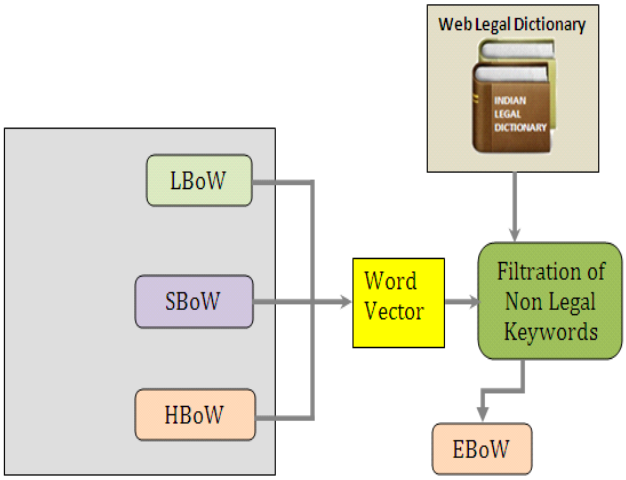

Fig. 1 Process to create Enriched Bag-of-Words

The process of preparing enriched Bag-of-Words (EBoW) for dowry act is a phased approach as shown in figure 1. In the first phase Legal Statute related to dowry act is preprocessed with java module [24] to generate words. This approach was earlier proposed by the same authors [24].The words are placed in a temporary vector which serves as a basic LBoW obtained from Legal Statute. In addition to this the SBoW derived from the earlier work of same authors [24] appended to this temporary vector. As an extension to this it is felt that human expertise is also valuable and enrich the legal repertoire, so the legal words suggested by an eminent Legal Advisor (Human Expert) appended to the earlier said temporary vector. The duplicate words from any of these above said three sources would automatically get pruned. From the perspective of the authors this would become a consummated vector containing the BoW for the dowry related legal issues. The effort of the authors is further refined with the help of word level comparison done with the Legal Dictionary [25] and the final filtration is done. Thus obtained vector contains unique words after the aforementioned process, hence it avoids duplicate words. In second phase vector elements are compared with external Legal Dictionary to filter out non legal keywords. After this process vector remains with pure legal keywords related to dowry act. Finally vector data converted into text file which is for future usage. The LBoW, HBoW and EBoW are presented in the annexure. The exclusive words which are available in EBoW and not present in the remaining other BoW's are as depicted in figure 2. The additional effort expended in terms of creating the exclusive words in the EBoW would be useful to the future researchers especially in the context of applying machine learning as well as text mining techniques. This could as well be extended to several different domains like science, social networking and medicine also.

Hence newly enriched BoW set consists more topic oriented terms and enhance the similarity measure process. Also improves the document classification process in legal domain.

Accomplice	affin	agre	alleg	amortization
Amend	amicus	ancillari	anti	alias
Anticipatory	arbiter	arraign	assult	attorney
Barrister	bar	beneficiari	betroth	bigami
Breach	bylaws	connubium	contravent	compensation
Conclusion	confession	conflict	contradict	concubin
Counter	culpabl	cozen	decre	demand
Digami	divorc	encourage	extend	espous
Family	famili	filiat	fingerpri nt	final
Foeticid	guardian	grounds	guilti	hang
Harm	high	injuri	insuffici	intent
Interrog	jail	juri	law	murder
Native	newlywed	murder	oblig	observe
Offenc	order	parti	penalti	petition
post-martum	prision	physic	reclaim	record
remand	respond	subjudice	statutori	threat
tribun	uxoricid	unlaw	voluntari	volit
vulner	wife	wit	writ	year
yield	zealou	zone	304B,498	406,498
256,498				

Fig. 2 Exclusive words present in EBoW

V. Comparitive Analysis & Results

For comparison of these four BoW's with Legal Dictionary words related to dowry issues by applying Cosine Similarity measure[3, 15] to identify word similarity across the BoW's generated. All the four BoW's and Dictionary Bow are converted into text files for finding similarity measure. A java program developed in order to run similarity measures on text files. The program accepts two text files as input and applies Cosine Similarity method on them. The output of the program generates score of similarity which lies between interval (0, 1). The four BoWs are compared with Legal Dictionary to estimate similarity of words. The similarity score results shown in Table 1 as given below.

SBoW	***HBoW***	***LBoW***	***EBoW***
0.3269 (32.69%)	0.2766 (27.66%)	0.4148 (41.48%)	0.6967 (69.67%)

Table 1. Cosine Similarity with Legal Dictionary

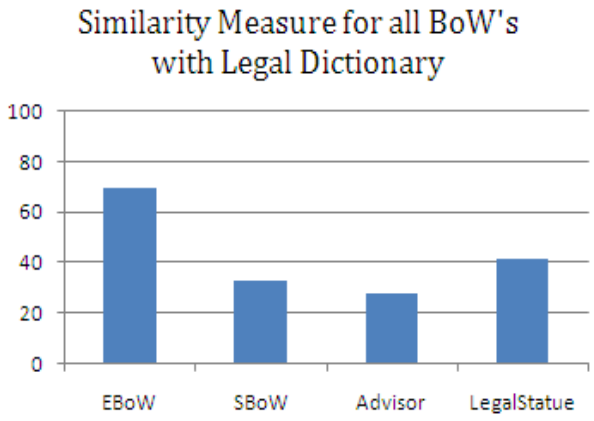

Figure 2. Word Similarity with Legal Dictionary

The results are as shown in figure 3. It is observed that EBoW has the highest similarity when compared with Legal Dictionary because of the reason that it is a composition of the remaining three. Among the independent BoW's Legal Statute has the greater similarity than the other two.

VI. Conclusion

The use of external knowledge sources such as Legal statute on dowry sections and Legal expert increased the word essence in extracted words. This enhancement improved Bag-of-Words granularity and reduced probability of non legal terms in words. Results shows that EBoW has the higher similarity measure in compared with other BoW's followed by LBoW. Better performance may be obtained by applying text mining techniques with EBoW. There is a further scope to enrich the BoW which would enhance word similarity by considering more number of external resources like Wikipedia and Word Net.

References

[1]. Eman Isamil and Walaa Gad, "CBER: An Effective Classification Approach Based on Enrichment Representation for Short Text Documents", University Press, Egypt, 2016.
[2]. Martin Vita, "Computing Semantic Textual Similarity based on Partial Textual Entailment", NLP Centre, 2015.
[3]. Prof. Sarika N Zaware, Mr. Asmit Gautam, Ms. Sumedha Nashte, Ms. Puneet Khanuja, "AN EFFECTUAL APPROACH FOR CALCULATING COSINE SIMILARITY", IJAERD, Vol.2, Issue 4, 2015.
[4]. Quoc Le, Tomas Mikolov,"Distributed Representations of Sentence and Documents", ICML conference, China, 2014.
[5]. S. Mizzaro, M. Pavan, I. Scagnetto and M. Valenti, "Short text Categorization Exploiting Contextual Enrichment and External Knowledge", Conference SoMeRa-14, ACM 978-1-4503-3022-0/14/07, 2014.
[6]. Peng Wang, C.Liu, H. Hao and Heng Zhang,"Short text Feature Enrichment Using Link Analysis on Topic-Keyword Graph", NLPC Conference, Springer, pp:79-90, 2014.
[7]. Muyu Zhang, B. Qin, Ting Liu and M. Zheng, "Triple based Background Knowledge Ranking for Document Enrichment", COLING Proceedings, pp: 917-927, Ireland, 2014.
[8]. R. Alfred, P. Anthony, S. Alias, Chim Kin O and L. H. Keng, "Enrichment of BOW Representation with Synatactic and Semantic Background Knowledge", M-CAIT conference, Springer, pp:283-292, 2013.
[9]. A. Ranjan Pal, A. Kundu, A. Singh, Raj Shekhar and K. Sinha, "An Approach to Word Sense Disambiguation Combining Modified LESK and Bag-of-Words", Conference (PDCTA), pp. 517-524, 2013.
[10]. Erich Schweighofer, Anton Geist, "Legal Query Expansion using Ontologies and Relevance Feedback", University Press, Vienna, Austria, 2013.
[11]. Yak Chu Li and H. M. Len, "Document Expansion using a side collection for Monolingual and Cross-language spoken Document Retrieval", ICSA Workshop, pp: 85-90, 2003
[12]. H.Duk Kim, Dae Hun Park and Ye Lu, "Enriching Text Representation with frequent Pattern Mining for Probabilistic topic modeling", ASIST Conference, USA,2012.
[13]. C. Carpineto and G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval", ACM Journal Library, 2012.
[14]. Md. Rafi, Sundus Hassan, Md. Shabid Shaikh, "Content based Text Categorization using Wikitology", NU-FAST, Karachi, 2012.
[15]. Ankush Maind, Prof. Anil Deorankar and Dr. Prashant Chatur, "MEASUREMENT OF SEMANTIC SIMILARITY BETWEEN WORDS: A SURVEY", IJCSEIT, Vol.2, No.6, December 2012.
[16]. Z. Dai, A. Sun and Xu-Ying Liu, "CREST: Cluster-based Representation Enrichment for Short Text Classification", NTUDIRP project, Singapore, 2010.
[17]. D. Andrei, San Jose, C. Muller, "Information Extraction from Legal Documents using Linguistic Knowledge and Ontologies", CAPES-Project, 2010.
[18]. "Termpedia for Interactive Document Enrichment", ACM Press, NLP-Research, Athens.
[19]. Hazram Imran and A. Sharan, "Thesaurus and Query Expansion", IJCSIT, Vol. 2, pp: 89-97, 2009.
[20]. L. Wang and Douglas W. Oard, "Query Expansion for Noisy Legal Documents", University of Maryland, University Press, 2007.
[21]. B. Biller beck and Justin Zobel, "Document Expansion versus Query Expansion for Ad-hoc Retrieval", ADCS Proceedings, Australia, 2005.

[22]. E. Gabrilovich and S. Markovitch, "Feature Generarion for Text Categorizatioon using World Knowledge", University Press, Technion, Israel, 2005.
[23]. E. Schweighofer, G. Haneder, A. Rauber and M. Dittenbach, "Improvement of Vector Representations of Legal Documents with Legal Ontologies", University Press, WIEN, Australia, 2002.
[24]. B.V.RamaKrishna, B.Basaveswar Rao, K. Chandan and K. Gangadhar rao, "An Enumerative Framework for Extraction of Bag-of-Words from Legal Documents", AJCSIT, ISSN: 2249-5126, pp: 62-66, Vol. 5, 2016.
[25]. http://www.supremecourtindia.com/Library/Legal Vocabulary/LegalDictionary.pdf

Appendix

304b,498	256,498	406,498	accus	act
articl	arrest	anti	appeal	bail
bench	blackmail	bodi	book	bribe
brutal	case	cell	charg	claus
complain	counsel	court	culprit	dead
death	deceas	dismiss	domest	dowri
dowry-death	evid	f.i.r	file	fine
forensic	govern	hang	harass	harm
high	hear	incedenc	imprison	insuffici
investigate	involuntari	judgement-no	judgment	kill
legal	legitim	legislate	liabl	magistrate
murder	regist	stai	petition	polic
property	public	punish	violence	witness
year				

Fig 3. Legal Bag-of-Words (LBoW)

abet	abuse	accompani	account	accused
act	administration	admission	admit	advocate
aid	alias	alimony	allegation	ambiguity
annul	accomplic	affen	alleg	anncilari
amend	amicus	amortization	anticipatory	appeal
appellant	approach	arbiter	arguendo	arrest
article	assualt	attempt	attest	attorney
audit	authority	arraign	bailee	bailor
bail	benficiari	betroth	bigami	bylaws
bar	barrister	basis	behavior	belief
bench	blackmail	body	book	breach
bribery	bride	brief	brought	brutal
burden	burn	connubium	case	cell
claim	commit	compensation	complaint	conclusion
condition	confession	conflict	court	crime
criminal	cruel	custody	death	deceased
contravent	counter	contradict	cozen	culabl
concubin	decre	digami	divorc	defense
demand	domestic	dowry	dowry-death	due
encourage	espous	extend	fingerrint	foeticid
famili	filiat	forensic	file	family
final	fine	force	gaurdian	girl
government	grounds	guilti	harassment	harm
harass	hang	helpless	high	higher
husband	injuri	insuffici	intent	imprison
interrog	include	injured	injuries	issued
judgment	judgment-no	jurisdiction	jail	juri
kill	law	legal	liability	local
lodge	magistrate	marriage	matter	murder
nativ	newlywed	notice oblig	offenc	order
parti	penalti	petition	post-martum	prision
petition	person	physic	prohibition	property

notice	observe	occurrence	offense	order
protect	petition	reclaim	remand	recognize
respondent	review	right	secure	session
statutori	sexual	show	significant	state
situation	social	specific	supreme	statute
stay	stayorder	subjudice	suicide	tribun
target	threat	threaten	tortured	type
unxoricid	unlaw	voluntari	volit	vulner
valuable	victim	violence	warrent	wife
woman	women	wit	writ	yield
young				

Fig 4. Enhanced Bag-of-Words (EBoW)

annul	accomplic	affin	agre	alleg
ancillari	arraign	beneficiari	betroth	bigami
connubium	contravent	counter	contradict	cozen
culabl	concubin	decre	demand	
digami	divorc	espous	famili	filiat
fingerrint	foeticid	gaurdian	guilti	hang
injuri	insuffici	intent	interrog	jail
juri	law	murder	nativ	oblig
offenc	order	parti	penalti	petition
post-martum	prision	reclaim	record	remand
stattutori	tribun	unxoricid	unlaw	voluntari
warrent	wife	wit	writ	year
yield	zone	zealou		

Fig 5. Legal Advisor Bag-of-Words (HBoW)