International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Effect of Technical Domains and Patent Structure on Patent Information Retrieval

Alok Khode, Sagar Jambhorkar

Abstract: ***Patents are critical intellectual assets for any competitive business. With ever increasing patent filings, effective patent prior art search has become an inevitably important task in patent retrieval which is a subfield of information retrieval (IR). The goal of the prior art search is to find and rank documents related to a query patent. Query formulation is a key step in prior art search in which patent structure is exploited to generate queries using various fields available in patent text. As patent encodes multiple technical domains, this work argues that technical domains and patent structure have their combined effect on the effectiveness of patent retrieval. The study uses international patent classification codes (IPC) to categorize query patents in eight technical domains and also explores eighteen different combination of patent fields to generate search queries. A total of 144 extensive retrieval experiments have been carried out using BM25 ranking algorithm. Retrieval performance is evaluated in terms of recall score of top 1000 records. Empirical results support our assumption. A two-way analysis of variance is also conducted to validate the hypotheses. The findings of this work may be helpful for patent information retrieval professionals to develop domain specific patent retrieval systems exploiting the patent structure.***

Keywords: Information Retrieval, Patent Retrieval, Prior Art, Query Formulation.

I. INTRODUCTION

In the increased global competition and knowledge-based economy, technological innovation is a determining factor of an organization's performance. At various stages of the technology life cycle, intellectual property rights (IPR) play a prominent role and thus, the demand for its protection has also increased. Patent is a key intellectual property right, providing maximum protection through exclusionary rights to inventor in exchange of public disclosure of an invention. The importance of patents is growing globally as it also promotes and encourages innovation. It is estimated that over 100 million patent documents exist worldwide and about 70% of the technological information disclosed in patents is never published anywhere else [1]. Hence, patent data can prove to be a gold mine, if it is retrieved, analyzed & utilized appropriately. Almost all R & D activity begins with a patent search as it helps avoiding reinventing the wheel and gives research directions. In this scenario, effective patent search is therefore considered as the crucial and primary task in patent informatics activities [2]. Prior art search is an important and the most common type of search in patent retrieval task. It involves retrieval of already available records similar to a given patent, henceforth referred to as a query patent or topic in this paper. Based on the prior art requirement, patent retrieval task can be referred by a variety of different names such as patentability search, novelty search, invalidity search and freedom-to-operate search [3], [4]. Prior art search helps patent examiners to check the novelty and validity of a claimed work [5]. Rise in the availability of patent related information and increase in the need to access this information, encourage researchers in the Information Retrieval (IR) domain to develop techniques for efficient and effective patent retrieval [6]. The retrieved information can be of help to various types of users like patent professionals, industrial and academic research communities.

Despite the remarkable advancements in the area of IR and search engine techniques, there still exists a gap between research in area of web searching and techniques usually adapted in patent retrieval [2]. This is due to the special nature of the patent document which makes patent retrieval a challenging task that requires robust and repeatable techniques/approaches [7], [8], [9], [10].

The challenges associated with patent retrieval are highlighted below:

- The whole patent document is used as a patent query. It is thus a defining challenge to build a query from a whole patent document which consists of various fields [11], [12], [13], [9], [10].
- Patents are multi-page, multi-topic, multi-modal, multi-language and semi-structured document. Patent retrieval is performed on diverse and large volumes of data [6].
- Use of vague, domain specific terminology, inventor specific non-standard acronyms/terms, presence of synonyms, homonyms and techno-legal words in patent document make the traditional keyword based search less effective [14], [15].
- Patent retrieval is a recall oriented task in which one cannot afford to miss a single relevant document while performing prior art search as it may lead to severe financial consequences due to lawsuit for patent infringements [16], [17].

Keeping the above challenges in mind, a lot of research in the patent domain is being carried out in order to improve the effectiveness of the prior art search.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Alok Khode*, Symbiosis Centre for Research and Innovation, Symbiosis International University, Pune, India. Email: alokkhode@gmail.com
Dr. Sagar Jambhorkar, Department of Computer Science, National Defense Academy, Pune, India. Email: sjambhorkar@yahoo.co.in

While carrying out such experiments, focus of the researchers is mostly on query formulations strategies exploiting the patent structure which consist of fields such as title, abstract, patent classification code, claims and description.

However, a need is felt to investigate the role of technical domains in the patent retrieval because a patent is confined to a single inventive concept which may cover multiple technical domains. Therefore the objective of this study is to investigate and evaluate the combine effect of technical domains and patent structure in formulating search strategies of patent retrieval. This may further give research directions to IR professionals to develop domain specific patent retrieval systems. This may also help patent professionals to exploit the patent structure by utilizing specific field or combination of fields available in patent text while working on particular subject domain.

In this work, international patent classification (IPC) code [18] is utilized to evaluate the effect of technical domains. The structure of the patent documents is exploited based on the previous researches in this area utilizing different fields or combination of fields of the patent document while formulating the query. Further, the documents are indexed, retrieved and ranked based on the BM25 function with default parameters using Terrier IR platform [19].

II. RELATED WORK

With the IR initiatives on patent retrieval such as NTCIR, TREC, CLEF-IP etc. and with the availability of domain specific benchmark dataset, patent retrieval has become an active area of research in information retrieval (IR) domain [20], [21]. Researchers have tested various techniques / methods such as query formulation, structured index, weighted fields, semantic search, document filtering, relevance feedback etc. for effective prior art search. Query formulation has been used as a preferred technique for better retrieval effectiveness. Usually, a whole patent is considered to be the query topic for prior art search. However, in traditional information retrieval tasks, the size of the topic is comparatively smaller which helps to increase the efficiency of the IR system as well as to minimize the possible noise in the result set. To achieve similar results in the patent retrieval, researchers have exploited patent structure consisting of various fields available in patent text [8]. Several post processing approaches have also been proposed to improve patent retrieval using patent classification codes (IPC) that identify patents by technical subject domain.

Previous works on query formulation have utilized query expansion to effectively capture user query intents [22], [23] and query reduction to negate the effect of ambiguous terms present in the huge patent queries [24]. Some of the early works on query formulation have used claims [25], [26] while few researchers have reported best performance over all other fields when description field and more specifically background summary of description field has been used as the source for extracting the query terms [12], [8]. However, Gobeill et al. [27] in their experiment suggest that use of description field degrades the performance as it contains more noise than information. Ganguly et al. [28] segmented the query based on various patent fields and then interleave the retrieval results of each of the query. Cetintas and Si [9] propose an automated patent retrieval approach by combining terms extracted from different fields and assigning them weights. Researchers have also utilized hierarchical structure of international patent classification (IPC) codes for retrieval as well as re-ranking and filtering at post-processing stage [29]. Mahdabi et al. [17] constructed a query specific patent lexicon from IPC definition and proposed proximity based framework. Giachanou et al. [30] have used patent classification codes to organize patents topic-wise by dividing the patent collection in to sub-collections and develop a method to effectively select relevant collection for the required prior art search. Graf et al. [13] explored the IPC codes to identify the similarity between patent documents which do not share similar terms. Mase et al. [31] propose a two stage patent retrieval method considering the claim structure. They used IPC to improve the recall while exploring claim structure to improve the precision.

Based on the literature review, it is observed that focus of the previous works is mostly on query formulation that exploits various metadata available in patent text. There is no work which attempts to know the interaction effect of technical domains and patent structure on the patent retrieval performance i.e. which particular field or combination of fields is important for effective retrieval when query belongs to a specific technical domain.

III. EXPERIMENTAL DESIGN

In this section, we first discuss about dataset and evaluation metric and then develop hypotheses for examining the effect of domain specific queries and combination of various patent fields on prior art search. We then explain methodologies for conducting our experiments. This study does not attempt to improve the overall effectiveness of the retrieval; rather different experiments are carried out to support the initial assumption mentioned in the introduction section.

A. Testing Collections and Pre-Processing

For the experiments, CLEF-IP 2011 datasets is used which consists of 3.5 million patent documents representing 1.5 million patents published by European Patent Office (EPO) and World Intellectual Property Organization (WIPO). The retrieval experiments described in this paper are implemented using Terrier which is a high performance and scalable information retrieval platform suitable for patent retrieval task. During indexing and retrieval, both documents and queries are stemmed using the porter stemmer. Patent specific stop-words available from US Patent and Trademark Office (USPTO) are removed as they appear so frequently in patent text that they lose their usefulness as search terms [32]. In the implementation, each section of a patent such as title, abstract, claims, IPC and description is indexed in a separate field.

However, when a query is processed, all indexed fields are targeted as this generally offers best retrieval performance [33]. The relevance assessment is provided for the topic set (queries) which are patents and have title, abstract, IPC, description, and claims sections. Only English topic subset which corresponds to 1351 patents is used. Total size of the dataset is approximately 20GB compressed data or 100GB uncompressed data.

B. Evaluation

Relevance judgments for the English language topics provided by CLEF-IP are used for evaluation. In general, various measures such as Recall, Precision and Mean Average Precision (MAP) are used to evaluate the performance of IR techniques.
MAP and precision give good and intuitive evaluation for IR task emphasizing precision. However, for the recall focused IR tasks such as patent retrieval, recall is a preferred metric for evaluation. Recall is the fraction of relevant documents that are retrieved. In this study recall@1000 (recall for top 1000 retrieved records) are reported while comparing and analyzing the result:

C. Hypotheses

Following three sets of hypotheses are defined:

1. Effect of technical domain (IPC sections):

H0: The mean recall of the IPC sections are equal
H1: The mean recall of the at least one IPC section is different

2. Effect of patent structure (patent field combinations):

H0: The mean recall of all field combinations are equal
H1: The mean recall of at least one field/field combination is different

3. Combined effect of technical domain and patent structure:

H0: There is no interaction between the field combinations and IPC sections for recall
H1: There is interaction between the field combinations and IPC sections for recall

To test the hypotheses, a two-way analysis of variance (ANOVA) is conducted using statistical tool SPSS 21 that examines the influence of technical domain and patent structure on the recall score of the patent retrieval. The technical domain is represented by IPC sections and the patent structure is represented by various patent fields or combination of fields.

D. Methodology

After indexing of patent fields, query formulation is considered as one of the major task in patent retrieval to fetch all the patents which are relevant to the query patent [8], [34]. In this paper, various experiments have been carried out to explore the possible permutations and combinations of available fields in patent text. Topic patents are clustered based on the IPC codes in order to evaluate the effect of the technical domains on the patent retrieval. IPC is a hierarchical classification system used to categorize patents according to the technical fields. This classification scheme contains more than 70,000 classification symbols and each patent is associated with one or more classification symbols. The IPC embodies 8 broad technical sections, labeled A through H. For example, a patent is assigned one IPC H03C 3/00. This is further illustrated in Fig. 1 and Fig. 2.

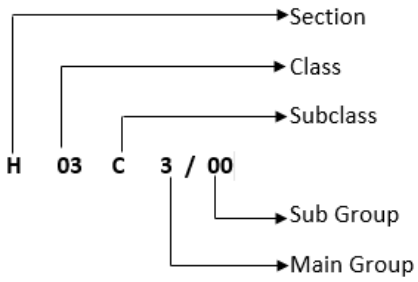

Fig. 1. Components of an IPC classification

H	Electricity
H03	Basic Electronic Circuitry
H03C	Modulation
H03C 3/00	Angle Modulation

Fig. 2. A hierarchical representation of the IPC

The distribution of these IPC sections over English topic set of CLEF-IP 2011 is given in the Table 1 below.

Table-1. IPC section distribution

IPC Section	Description	No. of patent topics (Multiple IPC section)	No. of patent topics (Exclusively single IPC section)
A	Human Necessities	250	190
B	Performing Operations; Transporting	213	128
C	Chemistry; Metallurgy	150	84
D	Textiles; Paper	23	15 (51*)
E	Fixed Constructions	18	10 (87*)
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting	143	91
G	Physics	263	153
H	Electricity	291	223

** Number of patent topics (exclusively with single IPC section) after populating claim and description fields from external sources*

It is suggested that in a standard IR evaluation approach, a set of minimum 50 queries is needed [35]. However, the number of query patents in IPC section D and section E were observed to be less. This is mainly because the claim and description fields for these IPCs were missing for most of the query patent and hence they were excluded from the topic subset of the CLEF-IP 2011 dataset.

To overcome this shortcoming, online patent databases such as Google-patents[36] and Espacenet[37] were used to populate missing information. For the experiment, a subset of 50 query patents is selected randomly for each of the IPC section where query patents belong to a single IPC section only.

Detail of all the fields, combination of fields used in the query formulation and their abbreviations is given in Table 2. For each of eight IPC specific topic sets, 18 different combinations of patent field combinations have been used to formulate queries. This implies to a total of 144 retrieval experiments which have been carried out.

The overall architecture of the retrieval system is illustrated in Fig. 3.

Probabilistic ranking function BM25 as suggested by Robertson et al. [38] is used for retrieving and scoring the documents with the default parameters (b = 0:75, k1 = 1:2, k3 = 1000) as BM25 scores are slightly more effective in practice for patent retrieval [39].

Given a query *Q*, containing keywords q1,q2…..qn, the BM25 score of a document *D* is:

$$BM25(Q,D) = \sum_{t \in Q \cap D} \left(\frac{tf_t^D \cdot log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot tf_t^Q}{tf_t^D + k_1 \cdot ((1-b) + b \cdot \frac{dl}{dl_{avg}})} \right)$$

Where tf^D_t is the frequency of term *t* in document *D*; *dl* is the document length; df_t is the document frequency of term *t*; dl_{avg} is the average document length in the entire collection; tf^Q_t is query term weight $=(k3+1)\ tf/k3+tf$; *tf* is the frequency of term *t* in query *Q*; *k1* and *k3* are term-frequency influence parameter and *b* is the document normalization influence parameter.

Table-2. Patent fields with their abbreviations

Sno	Patent field name	Abbreviation
1	Title	T
2	International Patent Classification (IPC)	I
3	Abstract	A
4	Claims	C
5	Full Description (DESC)	D
6	Description contains background of invention only (DBS)	Dbs
7	Title + Abstract	TA
8	Abstract + Claims	AC
9	Abstract + IPC	AI
10	Claim + IPC	CI
11	DESC + IPC	ID
12	DBS + IPC	IDbs
13	Abstract + DESC	AD
14	Abstract + DBS	ADbs
15	Title + Abstract + Claims	TAC
16	Title + Abstract + Claims + IPC	TACI
17	Title + Abstract + Claims + IPC + DESC	TACID
18	Title + Abstract + Claims + IPC + DBS	TACIDbs

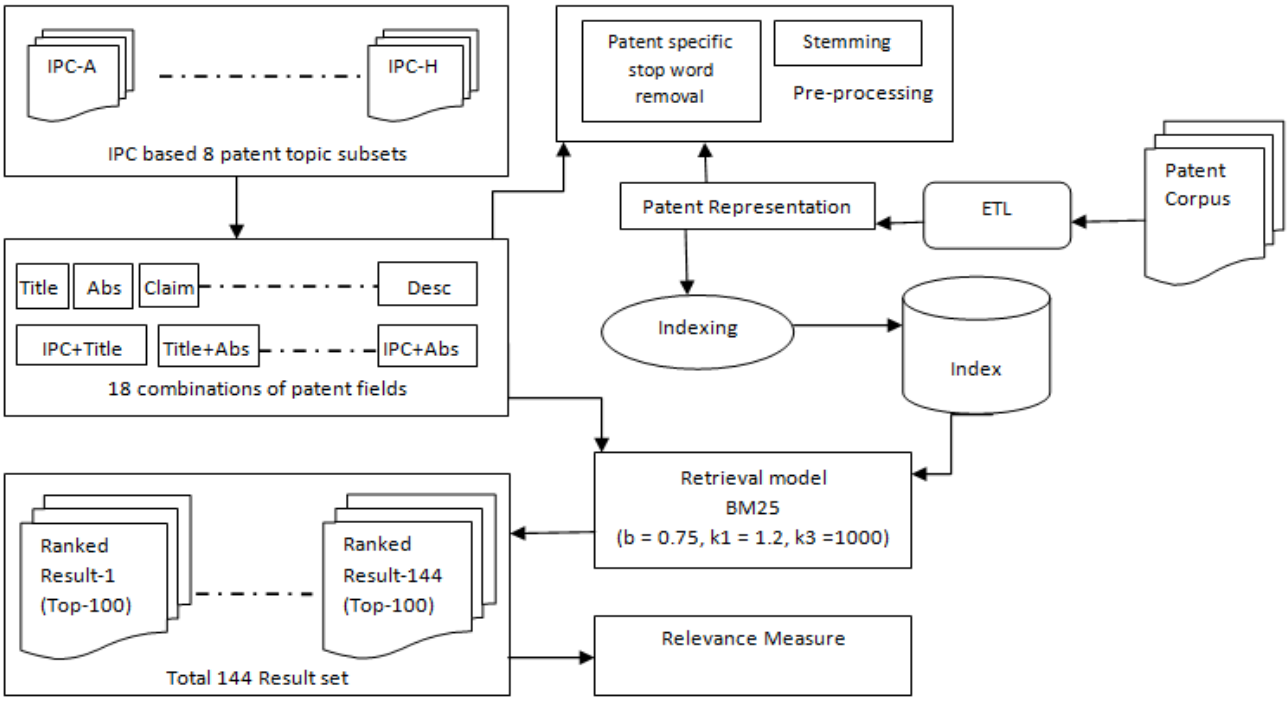

Fig. 3. Patent retrieval system architecture

IV. RESULT AND DISCUSSION

In this section, the results of the retrieval experiments are analyzed and discussed. A Factorial ANOVA is conducted to compare the main effects of two independent variables i.e. IPC sections and patent structure, and the interaction effect between these two on the recall metric of the patent retrieval experiments. To obtain the main effect and simple effects test, statistical tool SPSS 21 is used.

As given in Table 3, the recall score is analyzed with an 18X8, field combinations x IPC sections, factorial ANOVA. Each effect was tested with a MSE of .095. All the effects are statistically significant at the 0.05 significant level. The main effect for field combinations yield an F ratio of $F(17, 7{,}056) = 13.03$, $p < 0.001$ indicating significant different in the recall score between various field combinations. The main effect for IPC sections yield an F ratio of $F(7, 7{,}056)=63.07$, $p < 0.001$, indicating a significant difference in recall score between various IPC sections. Significant ($p \leq .05$) effects are also found for the main effect of field combinations X PC sections interaction, $F(119, 7{,}056) = 1.84$, $p <0.001$.

More interesting than the main effect of field combinations and IPC sections, however, is how the effect of various field combinations changed for difference IPC sections. As Pedhazur and Schmelkin[40] suggest that if interaction effects are present then the interpretation of main effect in isolation may be misleading. The significant interaction is further investigated by testing the simple main effects of IPC sections for each of the field combinations. Bonferroni adjustment is used for multiple comparisons and mean difference is considered at the .05 significant level. Considering the page restrictions, lengthy pairwise comparison table with more than 2000 rows generated due to 18X8 factorial ANOVA is not reported in this paper. However, this has been deposited in the Open Science Framework (OSF) repository (DOI 10.17605/OSF.IO/ZURAS) for reference purpose. Mean recall@1000 for various field combinations and IPC section is mentioned in Table 4.

Table-3. Tests of Between-Subjects Effects

Dependent Variable: Recall@1000					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	84.126a	143	.588	6.166	.000
Intercept	1699.471	1	1699.471	17,813.780	.000
field_combination	21.136	17	1.243	13.032	.000
ipc_section	42.121	7	6.017	63.073	.000
field_combination * ipc_section	20.868	119	.175	1.838	.000
Error	673.157	7,056	.095		

Total	2456.753	7,200			
Corrected Total	757.282	7,199			
a. R Squared = .111 (Adjusted R Squared = .093)					

Further investigation of significant difference between mean recall score ($p<0.05$) indicate that field 'I" yields significantly lower recall than any other field or combination of fields for IPC section A and IPC section C. All other field combinations give equal recall for these two IPC sections. In case of IPC section B, field combination 'ADbs', 'AI', 'IDbs' and 'TACIDbs' perform significantly better than 'I'. Interestingly for IPC section 'D', there is no statistically significant difference found in mean recall score amongst various fields or field combinations except 'AI' and 'T'. 'AI' gives better recall than 'T'. As far as IPC section 'E' is concern, an inverse pattern is on observed. 'I' performed better than 'A', 'AC', 'AD', 'ADbs', 'C', 'D', 'Dbs', 'ID', 'T', 'TAC' and 'AI' gives significantly better result than 'C', 'D' and 'T'. In case of IPC section 'F', combination 'AI' perform better than 'A', 'C' and 'T'. However, for other fields or combination of fields including 'I', there is no significant difference in terms of recall. Similar to most of the other IPC sections, 'I' gives lower recall as compared to other field combination except 'T' in IPC section 'G' and field 'T' gives lower recall than field combination 'AC', 'ADbs', 'AI', 'CI', 'Dbs', 'TAC', 'TACI' and 'TACIDbs'. However, there is no significant difference in mean recall between 'T' and 'I'. For IPC section 'H', mean recall is always lower as compared to 'AD', 'ADbs', 'CI', 'Dbs', 'ID', 'IDbs', 'TAC', 'TACI', 'TACID' and 'TACIDbs'. However, for all other combinations, there is no significant difference found for mean recall score.

Apart from the analysis based on two way anova, the investigation of the three best performing field combinations with respect to each technical domain discussed below may provide useful insights.

Table-4. Mean Recall@1000

	IPC Sections								
Field Combination	IPC_A	IPC_B	IPC_C	IPC_D	IPC_E	IPC_F	IPC_G	IPC_H	Marginal
A	0.539	0.438	0.559	0.411	0.28	0.432	0.56	0.375	0.449
AC	0.587	0.498	0.542	0.448	0.28	0.465	0.59	0.45	0.483
AD	0.606	0.496	0.569	0.438	0.274	0.478	0.564	0.48	0.488
ADbs	0.625	**0.549**	0.577	0.478	0.278	0.507	**0.606**	0.48	0.512
AI	0.616	**0.552**	**0.588**	**0.628**	**0.493**	**0.669**	**0.628**	0.421	0.574
C	0.588	0.503	0.539	0.419	0.254	0.442	0.581	0.452	0.472
CI	0.612	0.519	0.568	**0.523**	**0.385**	0.521	0.6	0.479	0.526
D	0.606	0.488	0.562	0.428	0.247	0.486	0.559	0.483	0.483
Dbs	0.614	0.524	0.558	0.457	0.272	0.519	0.589	0.481	0.502
I	0.261	0.31	0.191	0.443	**0.508**	**0.598**	0.228	0.236	0.347
ID	0.608	0.493	0.563	0.477	0.279	0.497	0.559	**0.485**	0.495
IDbs	**0.646**	**0.553**	0.569	**0.527**	0.339	**0.568**	0.6	**0.503**	0.538
T	0.492	0.351	0.439	0.378	0.236	0.41	0.364	0.285	0.369
TA	0.527	0.455	0.493	0.453	0.305	0.459	0.563	0.385	0.455
TAC	0.596	0.503	0.546	0.452	0.285	0.46	0.597	0.459	0.487
TACI	**0.644**	0.515	0.577	0.512	0.381	0.526	**0.607**	0.47	0.529
TACID	0.62	0.502	**0.581**	0.471	0.292	0.507	0.579	0.482	0.504
TACIDbs	**0.645**	0.54	**0.612**	0.5	0.338	0.522	0.6	**0.492**	0.531
Marginal	0.580	0.488	0.535	0.469	0.318	0.504	0.554	0.439	

The IPC section 'A' belongs to inventions related to human necessities consisting of areas such as agriculture, food-stuffs, personal/domestic articles, health care etc. It is observed that the combination of 'IDbs' yields best result. However, a slightly lower recall is achieved when combination 'TACIDbs' and 'TACI' are used. It is evident that patent classification proves to be an important field for effective query formulation, as it boost recall score for human necessities domain. Experiment also shows that rather than using lengthy description field to formulate the query, background of the invention (which is a part of broader description field) appears to be the best source for extracting terms for this technical domain.

The IPC B section belongs to performing various operations such as printing, transporting, shaping, nanotechnology etc. Result shows that the abstract and background of the inventions are the important fields while patent classification code helps to improve the recall score.

For IPC section C which covers chemistry and metallurgy related patents, experiment result favors the combination of all the fields i.e TACIDbs as source to formulate effective search query.

Inventions under IPC section D cover technologies related to textile and paper such as spinning, weaving, sewing, embroidery, paper-making etc. For this domain, combination of abstract and patent classification is a clear winner in terms of recall. Similar results is also observed for IPC section F and IPC section G. IPC F is related to mechanical engineering, lighting heating, weapons, blasting etc. while IPC G is meant for physics, instrumentation and information & communication technologies. Experimental result shows that for IPC section G, effective query can also be formulated when keywords are generated from 'TACI'.

Interestingly, for IPC section E which covers construction domains such as building of roads, railways, bridges, sewerages along with earth drilling & mining, only patent classification is more than sufficient to give comparatively higher recall.

As far as IPC section H is concerned which covers electrical domain, it is observed that combination 'IDbs' yields better result. However, a slightly lower recall is achieved when combination 'TACIDbs' and 'ID' are used. Hence it can be derived that role of title, claim and abstract is minimal while background of the invention along with patent classification outperform other field combinations.

It is also interesting to analyze the retrieval performance when a single field is used to formulate the query for each technical domain. As evident from Table 5, the patent classification code is not an effective field when it is used exclusively to formulate query across the domains except for IPC section E and IPC section F. This may be because the IPCs for these domains are well defined. In this experiment, patent classification has not been used at post processing for further filtering or ranking purpose. It is rather used as a supported field to yield better result. Our discussion with various IP search professionals reveal that IPC along with abstract, title and claims are the most preferred choice for query formulation across all the technical domains. However, our experiment shows that better results are achieved using various combinations of fields for different technical domains whereas the usage of single field is not that much effective. Out of all the single fields, background of the invention, which is a part of description field, performs better than any other fields. The description field as a whole is the lengthiest field comprising multiple aspects of the invention such as examples, tables, explanation of diagrams, definitions, technical background etc. This may be the reason for dilution of technical focus when query is generated from description field. However, background of the invention focuses on technical explanation of the invention, thus making it a better candidate for query formulation.

Table-5. Mean Recall@1000 for single patent field

	A	C	D	Dbs	I	T
IPC_A	0.539	0.588	0.606	**0.614**	0.261	0.492
IPC_B	0.438	0.503	0.488	**0.524**	0.31	0.351
IPC_C	0.559	0.539	**0.562**	0.558	0.191	0.439
IPC_D	0.411	0.419	0.428	**0.457**	0.443	0.378
IPC_E	0.28	0.254	0.247	0.272	**0.508**	0.236
IPC_F	0.432	0.442	0.486	0.519	**0.598**	0.41
IPC_G	0.56	0.581	0.559	**0.589**	0.228	0.364
IPC_H	0.375	0.452	**0.483**	0.481	0.236	0.285

Taking view of the above results, it becomes evident that there is interaction between technical domains, which are represented by patent classification codes, and patent structure, which consists of various fields. This proves our initial assumption that a single query formulation strategy cannot be effective across all the technical domains. For an effective patent retrieval system, subject specific search/retrieval techniques based on different query formulation settings need to be explored.

V. CONCLUSION AND FUTURE WORK

Finding similar patents is regarded as most important task in patent retrieval. Approaches proposed in the literature widely make use of query formulation as an effective technique for prior art search. In this study, different permutations and combinations of textual fields of patents have been used as a source to generate keywords for the query. Query patents have been categorized based on international patent classification codes to investigate the role of technical domain in the patent retrieval. Result of extensive experiments show that domain specific query formulation settings is needed in order to optimize the effectiveness of a patent retrieval. It is also established statistically that the subject domains and patent structure have their combined effect on patent retrieval. Use of various combinations of fields yield better results across different technical domains, whereas the usage of single field is not effective. Out of all the fields, background of the invention appears to be the best source for obtaining keywords while IPC may be supporting field to yield better results. In future work, further experiments would be required to find out effect of different settings of various IR techniques/algorithms on technical domains for effective retrieval. It would also be interesting to investigate the effect of IPC at post processing as well as effect of field based weighting methods and technical domains for patent retrieval.

REFERENCES

1. S. J. Wang, "The state of art patent search with an example of human vaccines", *Human vaccines*, vol. 7, no. 2, pp. 265-268, 2011.
2. W. Shalaby and W. Zadrozny, "Patent retrieval: a literature review", *Knowledge and Information Systems*, vol. 57, no. 153, pp. 1-30, 2019.
3. D. Hunt, L. Nguyen and M. Rodgers, Eds., *Patent searching: Tools & techniques*. Hoboken, NJ, USA: John Wiley & Sons, 2012.
4. D. Bonino, A. Ciaramella and F. Corno, "Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics", *World Patent Information*, vol. 32, no. 1, pp. 30-38, 2010.
5. S. A. Shaikh and B. R. Londhe, "Intricacies of software protection: a techno-legal review", *Journal of Intellectual Property Right*, vol. 21, no. 3, pp. 157 – 165, 2016.
6. Khode and S. Jambhorkar, "A Literature Review on Patent Information Retrieval Techniques", *Indian Journal of Science and Technology*, vol. 10, no. 3), pp. 1-13, 2017.
7. J. Tait and A. J. Trippe, *Current challenges in patent information retrieval* , vol. 29, M. Lupu and K. Mayer, Eds. Berlin, Germany: Springer, 2011.
8. P. Mahdabi, M. Keikha, S. Gerani, M. Landoni and F. Crestani, "Building queries for prior-art search", In *Information Retrieval Facility Conference*, Vienna, Austria, Jun. 2011, pp. 3-15.
9. S. Cetintas and L. Si, L. "Effective query generation and postprocessing strategies for prior art patent search" *Journal of the American Society for Information Science and Technology*, vol. 63, no. 3,pp. 512-527, 2012.

10. D. Zhou, M. Truran,J. Liu and S. Zhang, "Using multiple query representations in patent prior-art search", *Information retrieval*, vol. 17, no. 5-6, pp. 471-491, 2014.
11. X. Xue and W. B. Croft, "Automatic query generation for patent search", In *Proceedings of the 18th ACM conference on Information and knowledge management*, Nov. 2009, pp. 2037-2040.
12. X. Xue and W. B. Croft, "Transforming patents into prior-art queries", In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* , Jul. 2009, pp. 808-809.
13. E. Graf, I. Frommholz, M. Lalmas and K. Van Rijsbergen, "Knowledge modeling in prior art search", In *Information Retrieval Facility Conference*, Vienna, Austria, May 2010, pp. 31-46.
14. K. H. Atkinson, "Toward a more rational patent search paradigm", In *Proceedings of the 1st ACM workshop on Patent information retrieval*, Napa Valley, California, USA, Oct. 2008, pp. 37-40.
15. S. Bashir and A. Rauber, "Improving retrievability of patents in prior-art search", In *European Conference on Information Retrieval*, Milton Keynes, UK, Mar. 2010, pp. 457-470.
16. W. Magdy and G. J. Jones, "Examining the robustness of evaluation metrics for patent retrieval with incomplete relevance judgments", In *International conference of the cross-language evaluation forum for European languages*, Sep. 2010, pp. 82-93.
17. P. Mahdabi, S. Gerani,J. X. Huang and F. Crestani, "Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval", In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, Jul. 2013, pp. 113-122.
18. "*International Patent Classification (IPC)*", 2019. Accsseed on: Aug. 3, 2019. [Online]. Available: https://www.wipo.int/classifications/ipc/en/
19. "*Terrier IR Platform*". [Online]. Available: http://terrier.org. [Accessed: Jan. 3, 2019].
20. J. Tait, M. Lupu, H. Berger,G. Roda, M. Dittenbach, A. Pesenhofer,and C. J. Van Rijsbergen, "Patent search: An important new test bed for IR", In *Proc. 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009)*, Enschede, Netherlands, Feb. 2009, pp. 56-63.
21. L. Azzopardi, W. Vanderbauwhede and H. Joho, "Search system requirements of patent analysts", In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, Geneva, Switzerland, Jul. 2010, pp. 775-776.
22. F. Wang, L. Lin, S. Yang and X. Zhu, "A semantic query expansion-based patent retrieval approach", In *10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Shenyang, China, Jul. 2013, pp. 572-577.
23. P. Sharma, R. Tripathi and R. C. Tripathi, "Finding similar patents through semantic query expansion", *Procedia Computer Science*, vol. 54, pp. 390-395, 2015.
24. D. Ganguly, J. Leveling, W. Magdy and G. J. Jones, "Patent query reduction using pseudo relevance feedback", In *Proceedings of the 20th ACM international conference on Information and knowledge management*, Glasgow, UK, Oct. 2011, pp. 1953-1956.
25. K. Konishi, "Query Terms Extraction from Patent Document for Invalidity Search", In *Proceedings of NTCIR-5 Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, Dec. 2005. [Online]. Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-PATENT-KonishiK.pdf
26. P. Lopez and L. Romary. (2010). Experiments with citation mining and key-term extraction for prior art search. Presented at CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation, Padua, Italy. [Online]. Available: https://hal.inria.fr/inria-00510267.
27. J. Gobeill, E. Pasche,D. Teodoro and P. Ruch, "Simple pre and post processing strategies for patent searching in CLEF intellectual property track 2009", In *Workshop of the Cross-Language Evaluation Forum for European Languages*, Sep. 2009, pp. 444-451.
28. D. Ganguly, J. Leveling and G. J. Jones, "United we fall, divided we stand: A study of query segmentation and PRF for patent prior art search", In *Proceedings of the 4th workshop on Patent information retrieval*, Glasgow, Scotland, UK, Oct. 2011, pp. 13-18
29. B. Al-Shboul and S. H. Myaeng, "Query phrase expansion using wikipedia in patent class search", In *Asia Information Retrieval Symposium*, Dubai, UAE, Dec. 2011, pp. 115-126.
30. Giachanou, M. Salampasis and G. Paltoglou, "Multilayer source selection as a tool for supporting patent search and classification", *Information Retrieval Journal*, vol. 18, no. 6, pp. 559-585, 2015.
31. H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama and T. Oshio, "Proposal of two-stage patent retrieval method considering the claim structure", *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 4, no. 2, pp. 190-206, 2005.
32. "*Stopwords in full-text USTPO database*". [Online]. Available: http://patft.uspto.gov/netahtml/PTO/help/stopword.htm. [Accessed: Mar. 9, 2019].
33. M. R. Bouadjenek, S. Sanner and G. Ferraro, G. "A study of query reformulation for patent prior art search with partial patent applications", In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, San Diego, California, USA, Jun. 2015, pp. 23-32.
34. E. Pasche, J. Gobeill, O. Kreim, F. Oezdemir-Zaech, T. Vachon, C. Lovis and P. Ruch, "Development and tuning of an original search engine for patent libraries in medicinal chemistry. *BMC bioinformatics*, vol. 15, pp. 1, pp. S15, 2014.
35. C.D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2009.
36. "*Google Patents*". [Online]. Available: https://www.google.com/?tbm=pts. [Accessed: Dec. 5, 2018].
37. "*European Patent Office - Espacenet: patent database with over 100 million documents*".[Online]. Available: https://www.epo.org/searching-for-patents/technical/espacenet.html#tab-1. [Accessed: Nov. 5, 2018].
38. S. E. Robertson, S. Walker, M. Beaulieu, and P. Willett, "Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track", In *Proceedings of the 7th Text Retrieval Conference (TREC-7), NIST Special Publication SP, (500)*, 1999, pp. 253-264.
39. P. Mahdabi, L. Andersson, M. Keikha and F. Crestani, "Automatic refinement of patent queries using concept importance predictors", In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, Portland, USA , Aug. 2012, pp. 505-514.
40. E. J. Pedhazur and L. P. Schmelkin, *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers, 1991.

AUTHORS PROFILE

Alok Khode is a PhD research scholar at Symbiosis International University, Pune, India. He completed his Master of Computer Application from Devi Ahilya University, Indore, India in 1998. He has also been a Chevening Cyber Security Fellow with UK National Defence Academy and Cranfield University, UK. He is having more 20 years of software development and research experience with private and Government R & D organizations. His research interests are mainly but not limited to information retrieval and patent informatics.

Dr. Sagar Jambhorkar is an assistant professor in the Department of computer science at National Defence Academy, Pune, India. He received his PhD in computer science from Sant Gadge Baba University, Amravati, India. He is a member of IEEE, CSTA and IAEng. His research interests are mainly but not limited to image processing and information retrieval. His teaching interests include networking, artificial intelligence, software engineering, image processing, cyber security, cyber & Information warfare.