Strategies for Effective Chemical Information Retrieval

Suleyman Cetintas and Luo Si
Department of Computer Sciences, Purdue University
West Lafayette, IN, 47907, USA
{scetinta,lsi}@cs.purdue.edu

ABSTRACT

We participated in the technology survey and prior art search subtasks of the TREC 2009 Chemical IR Track. This paper describes the methods developed for these two tasks. For the technology survey task, we propose a method that constructs highly structured queries to do retrieval on different fields of chemical patents and documents in a weighted way. The proposed method i) enriches these structured queries with synonyms of the chemicals that have been identified, and ii) uses simple entity recognition to extract information for increasing or decreasing weights of some terms and to filter out documents from the ranked list. For prior art search task; we propose an automated query generation method that uses all title words, and selects sets of terms from the claims, abstract and description fields of query patents to transform a query patent into a search query. From the selected terms, chemical entities are extracted and synonyms for the identified chemical entities are included from PubChem. Then structured queries are formed to do retrieval over different fields of documents with different weights. Furthermore a post-processing step is also proposed that i) filters out some of the retrieved documents from the ranked list because of date constraints and ii) utilizes the IPC similarities between query patent and its retrieved patents to re-rank the retrieved documents. Empirical results demonstrate the effectiveness of these methods in both tasks.

1. INTRODUCTION

This paper describes the approaches used by members of Purdue University for technology survey and prior art search subtasks of the TREC 2009 Chemical IR Track. The Indri search engine[1] was utilized to index and retrieve various fields of documents, and its rich and powerful query language is exploited as it supports structured queries, handles synonyms, etc.

The test corpus used in this year's Chemical IR Track consists of 1,185,012 patent files from the chemical domain (classified under the IPC codes C and A61K), and covers patents in the field until 2007, registered at EPO, USPTO and WIPO (three major patent offices). The patents are in XML format, are provided by IRF[2] and contain title, claims fields along with description or abstract fields. Totally the uncompressed size of the patent files is 98.22GB. Along with chemical patent files, a total of 59,000 chemical journal articles (also in XML format) are also provided by the Royal Society of Chemistry[3], UK. The size of the set of scientific articles is approximately 3GB. Both of the sets of patent files and scientific articles are used for the technology survey task whereas only patent files are used for the prior art search task.

Domain specific information retrieval (IR) has recently been attracting more attention as important progresses have been made in IR in terms of theoretical models and evaluation. In addition to the Genomics and Legal tracks, Chemical IR Track has become another domain specific track of TREC and addresses the challenges generally in chemical IR and particularly in chemical patent IR. Although chemical IR can benefit the existing research in general purpose IR, there are distinct features in chemical IR that can be exploited. First of those distinct features is the structural information in the patents and articles. Despite a few exceptions [7], most prior research in the prior art search used the words from the claims field as the search query without examining other alternatives [2,3,4,6]. Although claims field is a very important field, other fields should also be carefully taken into account while selecting the terms for transforming patents into search queries in prior art search. In the same way, there is very limited research that also considers searching the queries in specific fields such as the abstract rather than in the whole documents [3]. Constructing a structured query by selecting query terms from various fields of documents and searching the constructed query over different fields of documents will be used as an approach in both technology survey and prior art search tasks in this work. The second distinct feature of chemical documents in general is that chemical

[1] http://www.lemurproject.org/indri/

[2] http://www.ir-facility.org/

[3] http://www.rsc.org/

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Strategies for Effective Chemical Information Retrieval	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Purdue University,Department of Computer Sciences,West Lafayette,IN,47907**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**6**	

```
#weight(
   4d #combine[titlegrp]( TERMS_FROM_QUERY_FIELD )
   4d #combine[invention-title]( TERMS_FROM_QUERY_FIELD )
   4d #combine[abstract]( TERMS_FROM_QUERY_FIELD )
   2d #combine[description]( TERMS_FROM_QUERY_FIELD )
   4d #combine[claims]( TERMS_FROM_QUERY_FIELD )
   2d #combine(TERMS_FROM_QUERY_FIELD )

   2d #combine[titlegrp]( TERMS_FROM_NARR_FIELD )
   2d #combine[invention-title]( TERMS_FROM_NARR_FIELD )
   2d #combine[abstract]( TERMS_FROM_NARR_FIELD )
   d #combine[description]( TERMS_FROM_NARR_FIELD )
   2d #combine[claims]( TERMS_FROM_NARR_FIELD )
   d #combine(TERMS_FROM_NARR_FIELD )
)
```

Table 1. A typical structured query (without entity detection enrichments) of the technology survey task that is searched over different fields of chemical patent and article documents in a weighted way. Note that most of the constructed technology survey queries are much more complex than this default query due to entity detection enrichments explained in section 2.1.5. TERMS_FROM_QUERY_FIELD is the set of all terms in the query field of a TS test topic and TERMS_FROM_NARR_FIELD is the set of all terms in the narrative (i.e. narr) field of a TS test topic. The weight d is chosen to be 1.

This query will be referred to as DEFAULT_TS_QUERY from now on.

molecules in those documents can be represented in multiple textual ways unlike other domains and a simple keyword search for a particular molecule using only one of its synonyms would retrieve only the documents with exact match and not the others. Therefore chemical molecules should be identified in the documents and synonyms of the identified molecules should be taken into account both for technology survey and prior art search tasks. The third distinct feature of chemical patent IR is that for prior art search the task is to find all relevant information (that may potentially invalidate the application patents claims of originality) published prior to the priority date of the application patent. The fourth distinct feature of chemical patent IR is the fact that unlike traditional IR where the precision of especially the top documents in the ranked list is very important, recall is more important in prior art search, since all relevant documents (within the date constraints) need to be retrieved. This is due to the fact that a single missed document can invalidate the query patent in prior art search. Last but not the least, all patents are assigned International Patent Classification (IPC) codes that can be exploited to calculate the similarity between a query patent and retrieved patents in prior art search.

The next section describes various approaches that utilize distinct features of chemical IR in detail.

2. SYSTEM DESCRIPTION

In this section, details of the proposed methods are described under two subsections, namely Query Construction Strategies and Post Processing Strategies.

2.1 Query Construction Strategies

This section describes the strategies that are used in both technology survey task and prior art search task for constructing the search queries.

2.1.1 Indexing

The Indri search engine was utilized to index the chemical patents and journal articles. To be able to do structured retrieval over different fields of patent and article files, Indri should be given the names of the particular fields that should be indexed. In this work, we indexed “titlegrp”, “invention-title”, “abstract”, “claims” and “description” fields in particular. We used Porter stemmer and removed the stopwords.

2.1.2 Feature Selection (Extraction of Query Terms)

For the technology survey task, all the terms in the title and description fields of the provided test topics are used. For prior art search task, the query itself is a patent file. So the search query should be automatically constructed from the query patent file. In particular, we use all the terms in the title field of the query patent file, top N terms with respect to a variant of TF-IDF scores (i.e. log(TF)*IDF) from the abstract, claims and description fields. N is chosen to be 30 in this work. Instead of selecting the terms from the whole patent files or only from a particular field (e.g. claims), we chose a particular number of

```
#weight(
      d #combine( SYNONYMS )

      2d #combine( TERMS_FROM_TITLE_FIELD )

      d #combine[abstract]( TERMS_FROM_ABST_FIELD )
      3d #combine( TERMS_FROM_ABST_FIELD )

      d #combine[claims]( TERMS_FROM_CLAIMS_FIELD )
      2d #combine(TERMS_FROM_CLAIMS_FIELD )

      d #combine[description]( TERMS_FROM_DESC_FIELD )
      2d #combine( TERMS_FROM_DESC_FIELD )
)
```

Table 2. A typical structured query of the prior art search task that is searched over different fields of chemical patent documents in a weighted way. SYNONYMS is the set of synonyms of identified chemicals, TERMS_FROM_TITLE_FIELD is the set of all terms in the title field, TERMS_FROM_ABST_FIELD is the set of selected terms from the abstract field, TERMS_FROM_CLAIMS_FIELD is the set of selected terms from the abstract field, and finally TERMS_FROM_DESC_FIELD is the set of selected terms from the description fields of a PA query patent. The weight d is chosen to be 1.

terms from each field to be able to have a better representation of the query patent in the constructed query file. Despite a few exceptions [7], prior approaches mostly used only the claims field for extracting the query terms [2,3,4,6]. Later when we do retrieval, we assign the weights of those terms accordingly. For example, the terms extracted from claims field will have a higher weight if they match some terms in the claims fields of the documents than the terms in other fields. This gives better similarity estimation during the retrieval between different fields of the query patent and the patents to be searched for in the prior art search task and will be explained more in Section 2.1.4.

2.1.3 Chemical Entity Recognition and Query Expansion with Synonyms from PubChem

A distinct feature of chemical documents in general is the fact that chemical molecules in those documents can be represented in multiple textual ways, and a simple keyword search would not suffice to have effective results. In this work, we extract the chemical entities in the (constructed) text queries by utilizing OSCAR3, an open source system that can identify much of the chemical terminology in chemical texts [1]. After the chemical entities are extracted, we include top 10 most commonly used synonyms of the identified chemicals from PubChem[4] in the query. Indri query language is utilized to integrate the synonyms of all identified chemicals into the automatically constructed queries with its powerful capabilities (using the {} operator) to handle synonyms of identified chemical entities.

2.1.4 Structured Retrieval over Chemical Patents and Articles

Chemical patents and articles are structured documents and the rich distinct information coming from structured nature of these documents can be exploited in both technology survey and prior art search. There is very limited prior research on searching the queries on different fields of documents in patent search literature [3]. In both technology survey and prior art search tasks of this work, we search our queries over different fields of the documents in a weighted way. In particular, using the Indri query language we construct a typical structured technology survey query as shown in Table 1. We basically i) give more importance to title, abstract and claims fields, but also consider the description field as well as the whole document; and ii) assign more weights to the query terms extracted from the query field of a TS test topic than the query terms extracted from the narrative field. In the same way, we construct a typical structured prior art search query as shown in Table 2. The approach is to search i) all query terms as well as the synonyms in the whole document and ii) query terms extracted from individual fields (i.e. claims, abstract, description) also in their corresponding fields. We don't search the terms extracted from the title field of the documents in the title fields since a typical title is too short to be effective for searching. The main intuition of this approach is that since query terms are also extracted from query patents which are also patent files, there may be more similarity between the same fields of a query patent and another patent that may potentially invalidate the query patent.

[4] http://www.ncbi.nlm.nih.gov/sites/entrez?db=pccompound

```
#combine(
    #scoreif(
        #uw( TERMS_FROM_TITLE_FIELD )

        DEFAULT_TS_QUERY
    )

    DEFAULT_TS_QUERY
)
```

Table 3. The technology survey structured query when the number of terms in the title are less than 3. The sub-query with the #scoreif operator first filters the documents that match the title and ignores the rest, then applies the default query over the filtered documents that match the title. The second sub-query performs default search. The #combine operator combines the scores of those two queries. We use the second sub-query here as a back-up when the first sub-query is too strict on filtering. So the first sub-query is expected to give high precision and the second sub-query is expected to give high recall. The combined query is a tradeoff in between.

2.1.5 Rule Based Entity Detection and Enrichment of Structured Queries

Entity detection techniques have been applied in various domains. In this work, we apply simple entity detection to extract valuable information that is later used to enrich the structured technology survey queries accordingly. Particularly the following manual rules were applied and the corresponding changes were done:

i. If the title of the query is 3 terms long or less, we treat those terms as very important. In particular, we use a query that is the combination of a default query shown in Table 1 and a query that first filters the documents that match the title and ignores the rest, then applies the default query over the filtered documents that match the title. Indri constructed version of such a query and more explanation can be found in Table 3.

ii. If there is a chemical that is identified in the title, do the same combination in i) but only with the identified chemical instead of the whole title.

iii. If there is an expression as "the use of", the words that come after this expression have higher importance and included in the DEFAULT_TS_QUERY in the same way the TERMS_FROM_NARR_FIELD are included. So the default query in this case has three groups of term sets (i.e. this new term set added to the existing two sets).

iv. If there is the term "not" without any auxiliary verb preceding it, then the terms following it have a negative meaning for the searched query. So we try to eliminate the documents with those terms that are not wanted. In particular, we construct a query similar to the one in Table 3, but we have only the first sub-query and the operator is #scoreifnot.

v. In chemical texts we often have expressions like "'chemical name' used as 'usage'" or "'chemical name' as 'usage'", therefore we utilize such "as" terms in the queries. If there are such uses of "as", i.e. if there is the expression " used as" or "'chemical name' as" in a sentence, then the terms following it are probably some specific uses of a chemical. Those terms are treated as the terms described in iii).

vi. If there is an expression as "the exact term", the words that come after this expression are treated as the terms in iii) and i). So we apply both approaches of incrementing the importance of those terms by applying strict filtering (to achieve high precision) in a combined way with a default query (to balance the recall) as explained in Table 3.

vii. If there is a date in the title or narrative, and there is the expressions "after, before, since, in, until" before the date; then we do date filtering after the retrieval, filtering out the results that are not relevant.

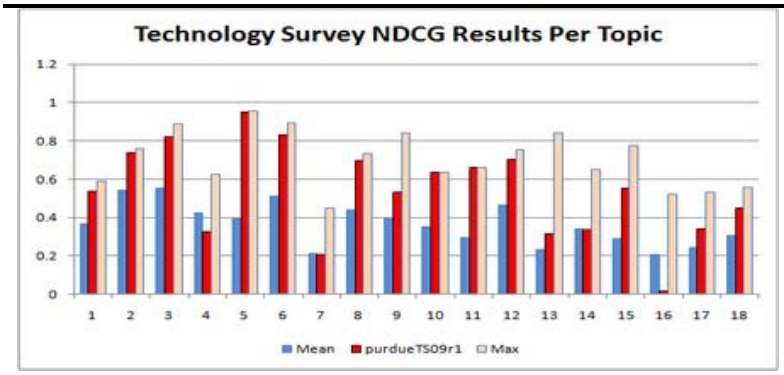
viii. If there are expressions describing the type of the source that is wanted, we also take into account those to filter only the desired document types. In particular, if there are terms such as "patents", "articles", "literature", "document" we check whether there is only one type of term. If the query mentions about only one source type, then only the documents in those type are returned.

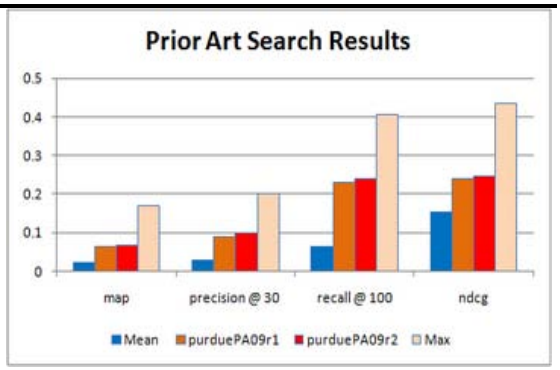
2.2 Post Processing Strategies

This section describes the post-processing strategies that are used in prior art search task for constructing the search queries.

2.2.1 Date Filtering on the Prior Art Search

In prior art search task, retrieved patents (that are expected to potentially invalidate a query patent) can be published before or after the query patent. Therefore some of the retrieved patents cannot invalidate the query patent as they may be published after the query patent: so doesn't violate the originality of the query patent. In this work, we discard the retrieved patents whose earliest priority dates are after the latest priority date of the query patent. If retrieved patent doesn't have priority dates, we use its publication date for comparison.





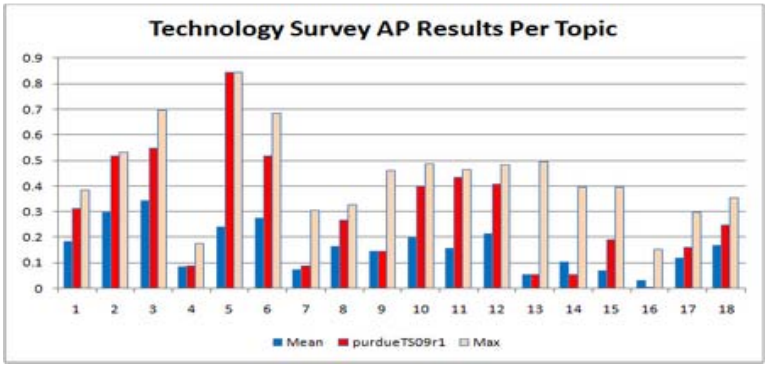


Table 4. Technology Survey Task NDCG and AP Results of purdueTS09r1 run with respect to Mean and Max scores (on the left) and Prior Art Search Task MAP, P@30, Recall@100, NDCG results of purduePA09r1 and purduePA09r2 runs with respect to Mean and Max scores (above). Note that prior art search task is a recall-oriented task.

2.2.2 Re-ranking Based on IPC Code Similarities

A distinct property of patent files is that all patents are assigned International Patent Classification (IPC) codes that can be exploited to calculate the similarity between a query patent and retrieved patents in prior art search. Prior research utilized the integration of IPC code similarity between a query patent and retrieved patents to re-rank the results in the prior art search literature [4,5]. Konishi compared the IPC codes of the query patent and retrieved patents [5]. If the retrieved patents have one or more IPC classes in common with the query patent, he multiplied the retrieval score by some constant [5]. Itoh used two approaches: in the first approach, he used the first 4 characters of the main IPC code (positioned at the first place of IPC description) of the query patent as a constraint over the retrieved documents; and in the second approach, he used the first 6 characters of the main IPC codes of the top 5 patents in the retrieved patents and used those IPC codes as a constraint over a baseline run (i.e. eliminated all retrieved patents that does not have any of those partial IPC codes) [4]. In this work, we used two features from the IPC code similarity: first 4 characters of the IPC code and first 11 characters of the IPC code. First four characters of the IPC code include section symbol, class number and subclass letter; and first 11 characters (including spaces) additionally include 1 to 3 digit "group" number, an oblique stroke and a number of at least two digits representing a "main group" or "subgroup". IPC eighth edition has a total of 8 sections, 129 classes, 639 subclasses, 7314 main groups and 61397 subgroups. The intuition behind using both first four characters and first 11 characters as a feature is to balance the tradeoff between precision and recall. The similarity calculated using the first 11 characters give high precision but is harder to achieve in most cases that leads to low recall; whereas the similarity calculated using the first 4 characters gives low precision (lots of similar patents in the retrieved patents) but gives high recall. In particular, the IPC code similarity between a query patent QP_i and a retrieved patent RP_j using the first 4 characters (i.e. $IPC^4Sim(QP_i, RP_j)$) is calculated as follows:

$$IPC^4Sim(QP_i, RP_j) = \frac{\sum_{n=1}^{|S^4_{QP_i}|}\sum_{m=1}^{|S^4_{RP_j}|}\delta(S^4_{QP_i}(n) == S^4_{RP_j}(m))}{|S^4_{QP_i}|} \quad (1)$$

where S^4_{QPi} is the set of partial IPC codes (i.e. first 4 characters) of a query patent QP_i, similarly S^4_{RPj} is the set of partial (first 4 characters of) IPC codes of a retrieved patent RP_j, $|S^4_{QPi}|$ is the number of unique partial IPC codes of QP_i, similarly $|S^4_{RPj}|$ is the number of unique partial IPC codes of RP_j, δ is the indicator function that returns 1 if the two compared IPC codes are the

same and 0 otherwise. The IPC code similarity between query patent QP_i and a retrieved patent RP_j using the first 11 characters (i.e. $IPC^{11}Sim(QP_i, RP_j)$) is calculated in a similar way.

After learning the two IPC code similarity features (i.e. $IPC^{4}Sim(QP_i, RP_j)$ and $IPC^{11}Sim(QP_i, RP_j)$), the retrieval score between QP_i and RP_j (i.e. $RetScore^{old}(QP_i, RP_j)$) is updated in a linear way as follows:

$$RetScore^{new}(QP_i, RP_j) = RetScore^{old}(QP_i, RP_j)\left(1 - \alpha(\lambda * IPC^{4}Sim(QP_i, RP_j) + (1-\lambda)IPC^{11}Sim(QP_i, RP_j))\right) \quad (2)$$

where α is a constant that controls the effect of IPC code similarity on the updated retrieval score and λ is a constant that controls the relative effect of $IPC^{4}Sim(QP_i, RP_j)$ and $IPC^{11}Sim(QP_i, RP_j)$ over the overall IPC similarity score. In this work α is set to 0.75 and λ is set to 0.2 (note that $RetScore^{old}(QP_i, RP_j)$ has a negative value).

3. EVALUATION

We submitted 1 run for technology survey task (purdueTS09r1) and 2 runs for the prior art search task (purduePA09r1 which is the mandatory run that was required by TREC Chemical IR track from all participants, and purduePA09r2) using our automatically constructed queries for all of them. Table 4 shows the performance of purdueTS09r1 run compared with the best and mean performance for technology survey task as well as the performance of purduePA09r1 and purduePA09r2 runs compared with the best and mean performance for prior art search task. Note that purdueTS09r1 run achieves the best (average across all topics) NDCG score across all submissions for the technology survey task and purduePA09r2 run achieves +264.6% improvement over the mean recall@100 score across all submissions for the prior art search task.

4. CONCLUSION

In this paper we describe the methods that we have developed for the technology survey and prior art search tasks of TREC 2009 Chemical IR Track. We studied various approaches for both tasks. In particular, for the technology survey tasks, we utilized structured retrieval, query expansion with synonyms of the detected chemical entities, rule based entity detection and filtering techniques. For prior art search task, we used feature selection (to select the query terms for transforming a query patent into a search query), structured retrieval, query expansion with synonyms of the detected chemical entities, date filtering and re-ranking of the results by utilizing IPC code similarities. Both of our approaches have an acceptable performance but still leave room for improvement.

5. ACKNOWLEDGEMENTS

Suleyman Cetintas and Luo Si have been supported by a research grant from National Science Foundation (IIS-0746830), a research grant from Indiana Economic Development Corporation, and a research grant from Purdue University.

6. REFERENCES

[1] Corbett, P. and Murray-Rust, P. 2006. High-throughput identification of chemistry in life science texts. *CompLife,* LNBI 4216:107–118.

[2] Fujii, A. 2007. Enhancing patent retrieval by citation analysis. In *Proceedings of the 30th Annual ACM SIGIR Conference on Research and Development on Information Retrieval, 599-606.*

[3] Itoh, H. 2004. NTCIR-4 patent retrieval experiments at ricoh. In *Proceedings of NTCIR Workshop 4 Meeting*.

[4] Itoh, H. 2005. NTCIR-5 patent retrieval experiments at ricoh. In *Proceedings of NTCIR Workshop 5 Meeting*.

[5] Konishi, K. 2005. Query terms extraction from patent document for invalidity search. In *Proceedings of NTCIR Workshop 5 Meeting, 312-317.*

[6] Mase, T.M., Ogawa, Y. 2005. Proposal of two-stage patent retrieval method considering the claim structure. ACM Transactions on Asian Language Information Processing 4(2) 186-202.

[7] Xue, X. and Croft, B. 2009. Transforming Patents into Prior-Art Queries. In *Proceedings of the 32nd Annual ACM SIGIR Conference on Research and Development on Information Retrieval, 808-809.*