Experience with and Suggestions for the Prior Art task of TREC Chemical (patent) track

Le Zhao and Jamie Callan
Language Technologies Institute, SCS
Carnegie Mellon University
2009/11/17
Count of hands

• Who receives emails from the TREC Chemical track mail-list?
• Who have seen a patent before the track?
• Who have looked at how patents cite other patents? Or invalidate other patents?
Prior Art task

• Task: Patent as query, Citations as relevant results
• Our approach
  – Date filtering (Prior)  [Aleksandr Belinskiy, Mail-list Comm]
    • Query patent:
      – Multiple priority dates – use latest priority date
    • Result patent:
      – Multiple dates – use publication date
  – Weighted bag of word queries (Relevant Art)
    • Title + Claims
    • Description
      – Too long, only used to weight terms
Indri Query Example

- `#filrej(#dateafter(07/07/1994))`
  - `#weight(0.6 #combine(detergent compositions))`
    - `0.4 #weight(16 1 14 bleaching 12 agent 11 composition 11 oxygen 11 7 10 4 8 u 8 2 7 o 7 available 6 claims 5 triazacyclononane 5 silver 5 coating 5 clo 5 organic 5 3 4 mn 4 minutes 3 co 3 mniii 3 0 3 5 3 bispyridylamine 3 description 3 n 3 containing 3 described 3 releasing 3 method 2 mixtures 2 time 2 compound 2 mixture 2 dentate 2 remainder 2 rate 2 mniv 2 source 2 tri 2 making 2 sprayed 2 intimate 2 completely 2 oac 2 cl 2 trimethyl 2 selected 2 premixed 2 bleach 2 dispersing 2 compositions 2 pf 2 released 2 perchlorate 2 oil 2 10 2 di 2 group 2 methyl 2 release 2 non 2 cobalt 2 consisting 2 interval 2 process 2 paraffin 2 particles 2 present 1 claim 1 perhydrate 1 nh 1 salt 1 copper 1 total 1 corrosion 1 bispyridyl 1 chloride 1 bi 1 8 1 dry 1 measured 1 partially 1 mnivbipy 1 och 1 trisdipyridylamine 1 comprises 1 mnivn 1 isothiocyanato 1 ligands 1 combination 1 triglycerides 1 bis 1 amine 1 6 1 bipy 1 binuclear 1 pyridylamine 1 mniiimniv 1 relaxing 1 inorganic 1 mixed 1 precursor 1 iron 1 hydrogenated 1 peroxyacid 1 additional 1 inhibitor 1 tetra 1 tris 1 level 1 derivatives 1 provided 1 diglycerides 1 gluconate 1 mono 1 wholly 1 complexed 1 catalyst ) ) )`
Initial experience

• 15 test topics (for training)
  – All EP patents

• keyword (title+claim) retrieval
  – MAP: 0.0115

• Date filtering
  – MAP: 0.0442

• Adding description weights
  – MAP: 0.0481
US patent dataset from Univ. of Iowa

• 15 US patent topics
  – Used for training
• keyword (title+claim) retrieval
  – MAP: 0.0586
• Date filtering
  – MAP: 0.1083
• Adding description weights
  – MAP: 0.1309
Zoom in

• Why overall performance low?
• Difference between two test sets?
Per topic performance

- **EP**
  - MAP=0.0481
    - Topic 1 easy: AP=0.4152
    - 14 topics MAP=0.0218
      - many topics AP=0
- **US**
  - Many topics with AP below 0.2
Zoom in

• Why so many low performing topics?
EP topic 3 (false positives)

• Q: *Oxygen-releasing (controlled release) bleaching agent*, with a non-paraffin oil organic silver coating agent, and additional corrosion inhibitor compound

• **Focus: top ranked results (cited means relevant)**
  
  Relevance: [Title]: [Summary of invention]
  
  – NR: *Controlled release* laundry bleach product (+ 2 more others)
  – NR: *Bleach* activation: improved bleach *catalyst* for low temperatures
  – NR: *Accelerated* release laundry bleach product
  – R: Bleach activation: activated by a *catalytic* amount of a transition metal complex
  – R: Concentrated detergent powder compositions: a surfactant, a detergency builder, enzymes, a peroxygen compound bleach and a manganese complex as effective bleach *catalyst*
Learnings (false positives)

• [Summary of the Invention] Important field:
  – for non-expert inspection (Amazon MechTurk?)
  – Maybe also for automatic retrieval
  – How do experts review patent applications?

• Bag-of-Word will fail in many cases
  – Most false positives have reasonably relevant descriptions (from a non expert’s eye)

• The most important part of a query patent is its novel part
  – Typically a small part of the document
EP topic 3 (misses)

• Focus: misses
  – Cited in the content:
  – “other catalyst examples include EPxxxx, USxxxx ...
  – about an unimportant area of the patent
  – These mentions also include the returned relevant patents
Learnings (misses)

• Patents cite related prior patents
  – Many citations are
    • mentions of prior arts made by the query patent
      – about an unimportant part of the patent
    • If these are relevant, all false positives can be relevant
  – Will a patent cite other patents that may invalidate itself?
    • Mechanism to ensure that? Increased application fee for finding other relevant results?

• For evaluation: what to include as Relevant?
  – Use the whole original reference list?
  – or only use citations added by others?
  – Patents have well marked search reports
    • For EP: X, Y, for US: *, judged by patent offices
      – Chem track assigns higher relevance to these [Florina Piroi Mail-list Comm]
    • EP 1-6, only 6 has 4 XYs, but a lot more applicant citations
      – We need better test sets
Discussion

• Patent experts and IR researchers
  – What fields to look at when reviewing patent?
  – Is it Possible to use Amazon MechTurk to do relevance judgements?
  – What patents do people cite when writing patents?
  – What kinds of patents do patent officers cite when reviewing patents?
  – What patents to use as relevant for the Chemical track?
EP topic 2 (false positives)

• **Q: Partial Oxidation of Sewage Sludge**
  – R: Slurry fuel comprised of a heat treated, partially dewatered sludge with a particulate solid fuel and its method of manufacture
  – NR: Environmentally safe process for disposing of toxic inorganic CN-containing sludge (partial oxidation gas generator)
  – NR: Fuel composition comprised of heat-treated dewatered sewage sludge and a biocide-containing fuel oil
EP topic 2 (misses)

• Related work:
  – “USxxxx... USxxxx... However, none of the references use our partial oxidation process.”