Patci – a tool for identifying scientific articles cited by patents

Sneha Agarwal, Miles Lincoln, Haoyan Cai, Vetle I. Torvik

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Introduction

Scientific research increasingly drives innovation and development of new technologies, and patent-to-paper citations can be used to trace this diffusion of knowledge and measure these science-to-technology spillover effects . However, the so-called "non-patent citations" in USPTO records do not contain authoritative identifiers, nor do they adhere to a standard format. They are strings written in free-form, often much too free, which makes it harder to systematically identify the articles or pieces of work cited.

Here, we introduce ***Patci*** -- a tool that takes a citation string and probabilistically identifies matching records from a set of bibliographic databases. It's distinctive features:

- Currently permits matching to biomedical literature (21.5M PubMed records) and computing/information sciences literature (3.2M DBLP records).
- Uses a probabilistic model trained on USPTO records but works well for citations originating from outside the patenting sphere.
- The algorithm extracts and weighs several hundred predictive features and does not rely on punctuation as delimiters of fields.
- A match probability as attached to each source link ID (e.g., PMID) which permits setting application-appropriate level of match stringency and permits sensitivity analysis.
- All 16M citations listed in granted USPTO patents (1975-present) have been processed and is available as a separate dataset.

Prior Work: An Illustrative Example

- ParseCit and FreeCite – taggers (not matchers) that rely heavily on punctuation as delimiters. Their reported precision/recall in the 90ies% rules them out as pre-processors for matching.
- lens.org – matcher (not probabilistic), unpublished algorithm

A "Non-patent" citation from USPTO patent record 03930718.

```
Analysis...Acoutstic Surface Waves on One Mirror; Hunsinger
Applied Optics; ol. 10, No. 2; Feb. 71; pp. 390-395.
```

FreeCite:

```
<Author>2 No</Author> <Title> Analysis... Acoutstic Surface
Waves on One Mirror; Hunsinger Applied Optics</Title>
<Volume>71</Volume> <Pages>390-395</Pages>
```

Corresponding PubMed PMID: 20094456

```
<Author>Hunsinger BJ </Author> <Title>Analysis of an
interferometer with acoustic surface waves on one
mirror</Title> <Journal>Appl Opt</Journal>
<Year>1971</Year> <Month>Feb</Month> <Day>1</Day>
<Volume>71</Volume> <Issue>2</Issue> <Pages>390-5</Pages>
```

Patci matching (labels with gives frequencies)

```
Tanalysis(632575) TPacoutstic:Tacoutstic|Tacou*stic(20406)
Tsurface(122347) Twaves(13518) Tmirror(16839)
Ahunsinger(82) Japplied(483549) Joptics(94291) Xol(3103)
V10(1238768) N2(4288062) Mfeb(1458806) YP71:1971(557157)
P390(34652) Q395(42708)
A82 T1 J39606 V1238768 N4288062 P1371 M1458806 Y557157
JVNMY38 AO82 TO1
```

Training a probabilistic model

Scope

- Processed 16M citations from 1.7M granted USPTO patents
- Indexed 21.5M records from PubMed; 3.2M from DBLP but the infrastructure can easily plug-n-play others.

Training data – automatically generated

- Positive examples: high proportion (and 1st-ranked) in proportion of matching terms (with some adjustments)
- Negative examples: 1) low proportion (or 2nd ranked) of matching terms, 2) key words like "advertisement", "search report", etc.

Several hundred match features

- Alphanumeric field-specific term matching (A, B, C, T, J, V, N, M, Y, P, Q)
- Misspelling (e.g., edit-distance("Acoutstic","Acoustic") = 1)
- Variant spelling (e.g., journal acronyms like PNAS)
- Partial matching (e.g., page number within range),
- Term frequency (e.g., Hunsinger is rare: frq = 82)
- Matching field length and proportion (e.g., title matched 20 of 25 characters)
- Field order (e.g., Journal followed by Volume)
- Time (e.g., citation before publication?)

The prediction model

- Logistic regression
- Semi-automated attribute selection

Application of the trained model

1. Generate candidate matches ordered by proportion of exact matches (Sphinx text indexing of MySQL tables enables quick retrieval and provides term frequencies)
2. Calculate match probability for all candidates and rank
3. Store in database for fast future lookups

Evaluation

Sample generation

- Unequal probability sampling – includes more difficult/borderline cases (with probability<0.1 and probability>0.9)

Human feedback

- Each pair of citation and its potential match is evaluated twice (by two different people)

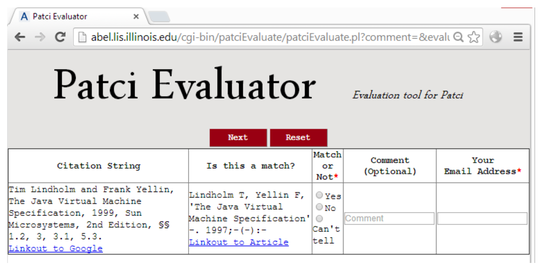

Fig.1: Screenshot of Patci Evaluator webpage used by evaluators.

Results

Evaluators often disagree … because

a) The task is difficult and time was limited
Oversampling cases where Patci was unsure.

b) Differing operating definitions of a match
Are we establishing links to the author(s), the work, the article, or a version of an article. Some might consider a technical report vs. the article later published in a journal not a match.

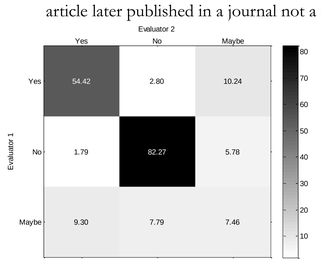

Fig.2: Confusion matrix of evaluators' responses (Jaccard distances shown).

Patci mimics a lenient matcher

- Lenient match = at least one evaluator says "Yes, a match"
- Strict match = both evaluators say "Yes, a match"
- Precision/recall estimated using Horvitz-Thomson normalization due to unequal probability sampling

Cut-off	0.9		0.5	
	Precision	Recall	Precision	Recall
Lenient	0.9996	0.9841	0.9951	0.9935
Strict	0.9951	0.9713	0.9309	0.9912

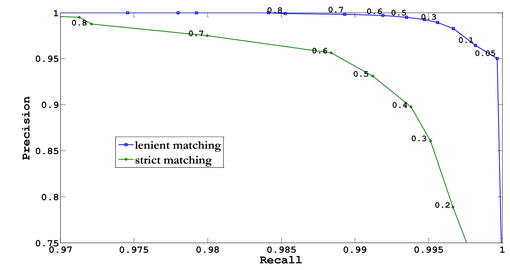

Fig.3: Precision vs. recall of Patci matching assuming evaluator(s) are correct. Match probability cut-offs are listed along the curves.

Citations to science has increased exponentially

- And the proportion to biomedical science has peaked

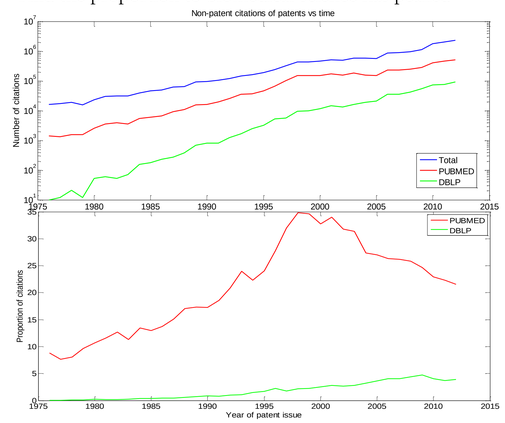

Fig4: Number of citations to PUBMED and DBLP articles have been increasing with time

Patci is remarkably efficient and accurate – typically less than a second per citation and often more accurate than humans.

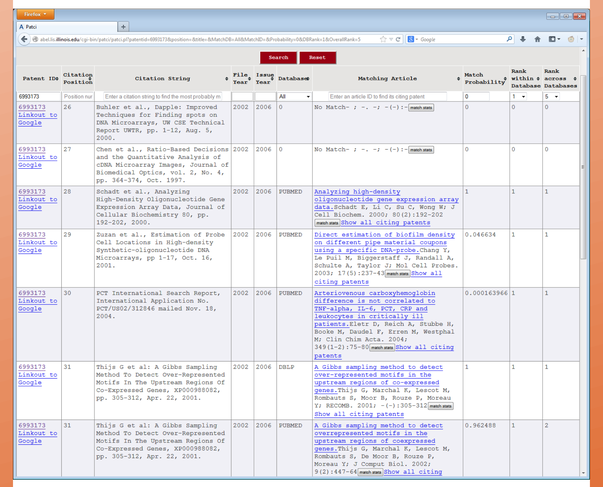

Fig.5: Patci screenshot when searching for USPTO patent number as input.

Conclusions and Planned Work

Patci, as a search tool or dataset, should be a useful resource for studies on the interactions between scholarly and industrial work

Challenges we still plan to address include:

- Fine-tuning for short nearly ambiguous citations
- Filtering for citations to other aggregators (e.g., Chemical Abstracts), foreign or non-granted patents and search reports, webpages, etc.
- Additional databases (e.g., books, dissertation, and other genres) and citation collections (e.g., citations from NSF grants)

Acknowledgments

Research reported in this publication was supported in part by the National Institute on Aging of the NIH (Award Number P01AG039347) and the Science of Science and Innovation Policy program of the NSF (Award Number 0965341). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF.

Many thanks to Brent Fegley and Shubhanshu Mishra for their insightful comments and valued feedback.

ILLINOIS