CLEF–IP 2009: retrieval experiments in the Intellectual Property domain

Giovanna Roda
Matrixware
Vienna, Austria

CLEF 2009 / 30 September - 2 October, 2009

Previous work on patent retrieval

CLEF-IP 2009 is the first track on patent retrieval at CLEF[1] (Cross Language Evaluation Forum).

[1] http://www.clef-campaign.org

Previous work on patent retrieval

CLEF-IP 2009 is the first track on patent retrieval at CLEF[1] (Cross Language Evaluation Forum).

Previous work on patent retrieval:

[1] http://www.clef-campaign.org

Previous work on patent retrieval

CLEF-IP 2009 is the first track on patent retrieval at CLEF[1] (Cross Language Evaluation Forum).

Previous work on patent retrieval:

- ACM SIGIR 2000 Workshop

[1] http://www.clef-campaign.org

Previous work on patent retrieval

CLEF-IP 2009 is the first track on patent retrieval at CLEF[1] (Cross Language Evaluation Forum).

Previous work on patent retrieval:

- ACM SIGIR 2000 Workshop
- NTCIR workshop series since 2001

[1] http://www.clef-campaign.org

Previous work on patent retrieval

CLEF-IP 2009 is the first track on patent retrieval at CLEF[1] (Cross Language Evaluation Forum).

Previous work on patent retrieval:

- ACM SIGIR 2000 Workshop
- NTCIR workshop series since 2001
  Primarily targeting Japanese patents.

[1] http://www.clef-campaign.org

Previous work on patent retrieval

CLEF-IP 2009 is the first track on patent retrieval at CLEF[1] (Cross Language Evaluation Forum).

Previous work on patent retrieval:

- ACM SIGIR 2000 Workshop
- NTCIR workshop series since 2001
  Primarily targeting Japanese patents.
  - ad-hoc task (goal: find patents on a given topic)
  - invalidity search (goal: find patents invalidating a given claim)
  - patent classification according to the F-term system

[1] http://www.clef-campaign.org

Legal and economic implications of patent search.

- patents are legal documents
- patent portfolios are assets for enterprises
- a single patent search can be worth several days of work

High recall searches

Missing even a single relevant document can have severe financial and economic impact. For example, when a granted patent becomes invalidated because of a document omitted at application time.

CLEF–IP 2009: the task

The main task in the CLEF–IP track was to find *prior art* for a given patent.

CLEF–IP 2009: the task

The main task in the CLEF–IP track was to find *prior art* for a given patent.

Prior art search

Prior art search consists in identifying all information (including non-patent literature) that might be relevant to a patent's claim of novelty.

Prior art search.

The most common type of patent search. It is performed at various stages of the patent life-cycle and with different intentions.

Prior art search.

The most common type of patent search. It is performed at various stages of the patent life-cycle and with different intentions.

- before filing a patent application (*novelty search* or *patentability search* to determine whether the invention fulfills the requirements of

Prior art search.

The most common type of patent search. It is performed at various stages of the patent life-cycle and with different intentions.

- before filing a patent application (*novelty search* or *patentability search* to determine whether the invention fulfills the requirements of
  - novelty

Prior art search.

The most common type of patent search. It is performed at various stages of the patent life-cycle and with different intentions.

- before filing a patent application (*novelty search* or *patentability search* to determine whether the invention fulfills the requirements of
  - novelty
  - inventive step

Prior art search.

The most common type of patent search. It is performed at various stages of the patent life-cycle and with different intentions.

- before filing a patent application (*novelty search* or *patentability search* to determine whether the invention fulfills the requirements of
  - novelty
  - inventive step
- before grant - results of search constitute the search report attached to patent document

Prior art search.

The most common type of patent search. It is performed at various stages of the patent life-cycle and with different intentions.

- before filing a patent application (*novelty search* or *patentability search* to determine whether the invention fulfills the requirements of
  - novelty
  - inventive step
- before grant - results of search constitute the search report attached to patent document
- *invalidity search*: post-grant search used to unveil prior art that invalidates a patent's claims of originality

The patent search problem

Some noteworthy facts about patent search:

The patent search problem

Some noteworthy facts about patent search:

- *patentese*: language used in patents is not natural

The patent search problem

Some noteworthy facts about patent search:

- *patentese*: language used in patents is not natural
- patents are linked (by citations, applicants, inventors, priorities, ...)

The patent search problem

Some noteworthy facts about patent search:

- *patentese*: language used in patents is not natural
- patents are linked (by citations, applicants, inventors, priorities, ...)
- available classification information (IPC, ECLA)

The patent search problem

Some noteworthy facts about patent search:

- *patentese*: language used in patents is not natural
- patents are linked (by citations, applicants, inventors, priorities, ...)
- available classification information (IPC, ECLA)

Outline

Outline

The CLEF–IP Patent Test Collection

The CLEF–IP collection comprises

- target data: 1.9 million patent documents pertaining to 1 million patents (75GB)

The CLEF–IP Patent Test Collection

The CLEF–IP collection comprises

- target data: 1.9 million patent documents pertaining to 1 million patents (75GB)
- 10,000 topics

The CLEF–IP Patent Test Collection

The CLEF–IP collection comprises

- target data: 1.9 million patent documents pertaining to 1 million patents (75GB)
- 10,000 topics
- relevance assessments (with an average of 6.23 relevant documents per topic)

The CLEF–IP Patent Test Collection

The CLEF–IP collection comprises

- target data: 1.9 million patent documents pertaining to 1 million patents (75GB)
- 10,000 topics
- relevance assessments (with an average of 6.23 relevant documents per topic)

Target data and topics are multi-lingual: they contain fields in English, German, and French.

Patent documents

The data was provided by Matrixware in a standardized XML format for patent data (the Alexandria XML scheme).

```
<?xml version="1.0" encoding="UTF-8" ?>
- <patent-document status="new" lang="EN" ucid="EP-0481532-B1"
  country="EP" doc-number="0481532" kind="B1" date="19951227">
  + <bibliographic-data>
  + <description lang="EN" status="new">
  + <claims lang="DE" status="new">
  + <claims lang="EN" status="new">
  + <claims lang="FR" status="new">
  + <legal-status status="new">
  </patent-document>
```

Looking at a patent document

```
<?xml version="1.0" encoding="UTF-8" ?>
- <patent-document status="new" lang="EN" ucid="EP-0481532-B1"
    country="EP" doc-number="0481532" kind="B1" date="19951227">
  + <bibliographic-data>
  - <description lang="EN" status="new">
      <p num="">The present invention relates to a semiconductor
      memory device, more particularly it relates to a nonvolatile
      memory device constituted by combining a volatile memory
      cell and a nonvolatile memory cell including a floating gate
      circuit element.</p>
      <p num="">Recently, in a static random access memory device
      (RAM), a volatile memory cell is combined with a floating
      gate circuit element to obtain a nonvolatile memory cell
      which is used to constitute a nonvolatile memory device. In a
      nonvolatile memory device of this type, the circuit
      configuration of each memory cell tends to be complex, and
      so the size of each memory cell tends to be large. However,
      this tendency leads to degradation in the reliability and
      integration of the memory device. In view of this problem,
```

Field: **description**

Language: German **English** French

Looking at a patent document

```
+ <bibliographic-data>
+ <description lang="EN" status="new">
- <claims lang="DE" status="new">
  + <claim num="xx">
  - <claim num="xx">
      <claim-text>Halbleiter-Speichervorrichtung nach Anspruch 1,
        bei der die Schreib-Schaltungsanordnung eine erste
        Schreibschaltung umfaßt, die mit dem anderen Anschluß
        des Tunnelkondensators (TC61 , TC62 , TC71 , TC81 , TC91 ,
        TCa) verbunden ist, und eine zweite Schreibschaltung
        aufweist, die mit dem einen Anschluß des
        Tunnelkondensators kapazitiv gekoppelt ist.</claim-text>
    </claim>
  - <claim num="xx">
      <claim-text>Halbleiter-Speichervorrichtung nach Anspruch 2,
        bei der die flüchtige Speicherzelle (1) zwei Transistoren
        (Q1 , Q2 ) aufweist, die zueinander kreuzgekoppelt sind,
        wobei die Spannungen der Gateanschlüsse der zwei
        Transistoren der nicht flüchtigen Speicherzelle (21) als
```

Field: **claims**

Language: **German** English French

Looking at a patent document

+ <bibliographic-data>
+ <description lang="**EN**" status="**new**">
+ <claims lang="**DE**" status="**new**">
- <claims lang="**EN**" status="**new**">
  - <claim num="**xx**">
    <claim-text>**A semiconductor memory device comprising a volatile memory cell (1, 11, 11') and a nonvolatile memory cell (16, 17, 20, 21, 22, 23) corresponding to the volatile memory cell, said nonvolatile memory cell comprising a transistor (Q_{51} , Q_{73} , Q_{81} , Q_{93} , TM) which has a floating gate and is operable to turn on or off in response to memorised data, a tunnel capacitor (TC_{51} , TC_{52} , TC_{71} , TC_{81} , TC_{91} , TCa), one electrode of which is connected to the floating gate, and write circuitry coupled to the tunnel capacitor, for supplying a high voltage (VH1, VH2; VH1, VH) as a write voltage to said tunnel capacitor in response to the memorised data of said volatile memory cell; characterised in that: said write circuitry includes first and second high voltage sources (VH1, VH2;**

Field: **claims**

Language: German **English** French

Looking at a patent document

+ <bibliographic-data>
+ <description lang="**EN**" status="**new**">
+ <claims lang="**DE**" status="**new**">
+ <claims lang="**EN**" status="**new**">
- <claims lang="**FR**" status="**new**">
  - <claim num="**xx**">
    <claim-text>**Dispositif de mémoire à semiconducteur comprenant une cellule de mémoire volatile (1, 11, 11') et une cellule de mémoire non volatile (16 ,17, 20, 21, 22, 23) correspondant à la cellule de mémoire volatile, ladite cellule de mémoire non volatile comprenant un transistor (Q_{51} , Q_{73} , Q_{81} , Q_{93} , TM) qui comporte une grille flottante et qui sert à activer ou désactiver, en réponse à des données mémorisées, un condensateur tunnel (TC_{51} , TC_{52} , TC_{71} , TC_{81} , TC_{91} , TCa) dont une électrode est connectée à la grille flottante, et un circuit d'écriture couplé au condensateur tunnel, pour appliquer une tension élevée (VH1, VH2; VH1, VH) en tant que tension d'écriture audit condensateur tunnel, en réponse aux données mémorisées de ladite cellule de mémoire**

Field: **claims**

Language: German English **French**

Topics

The task for the CLEF–IP track was to find prior art for a given patent.

Topics

The task for the CLEF–IP track was to find prior art for a given patent.
But:

Topics

The task for the CLEF–IP track was to find prior art for a given patent.
But:

- patents come in several versions corresponding to the different stages of the patent's life-cycle

Topics

The task for the CLEF–IP track was to find prior art for a given patent.
But:

- patents come in several versions corresponding to the different stages of the patent's life-cycle
- not all versions of a patent contain all fields

Topics

How to represent a patent topic?

Topics

We assembled a “virtual patent topic” file by

Topics

We assembled a "virtual patent topic" file by

- taking the B1 document (granted patent)

Topics

We assembled a “virtual patent topic” file by

- taking the B1 document (granted patent)
- adding missing fields from the most current document where they appeared

Criteria for topics selection

Patents to be used as topics were selected according to the following criteria:

1. availability of granted patent
2. full text description available
3. at least three citations
4. at least one highly relevant citation

Relevance assessments

Relevance assessments

We used patents cited as prior art as relevance assessments.

Relevance assessments

We used patents cited as prior art as relevance assessments.

Sources of citations:

Relevance assessments

We used patents cited as prior art as relevance assessments.

Sources of citations:

1. applicant's disclosure: the USPTO requires applicants to disclose all known relevant publications

Relevance assessments

We used patents cited as prior art as relevance assessments.

Sources of citations:

1. applicant's disclosure: the USPTO requires applicants to disclose all known relevant publications
2. patent office search report: each patent office will do a search for prior art to judge the novelty of a patent

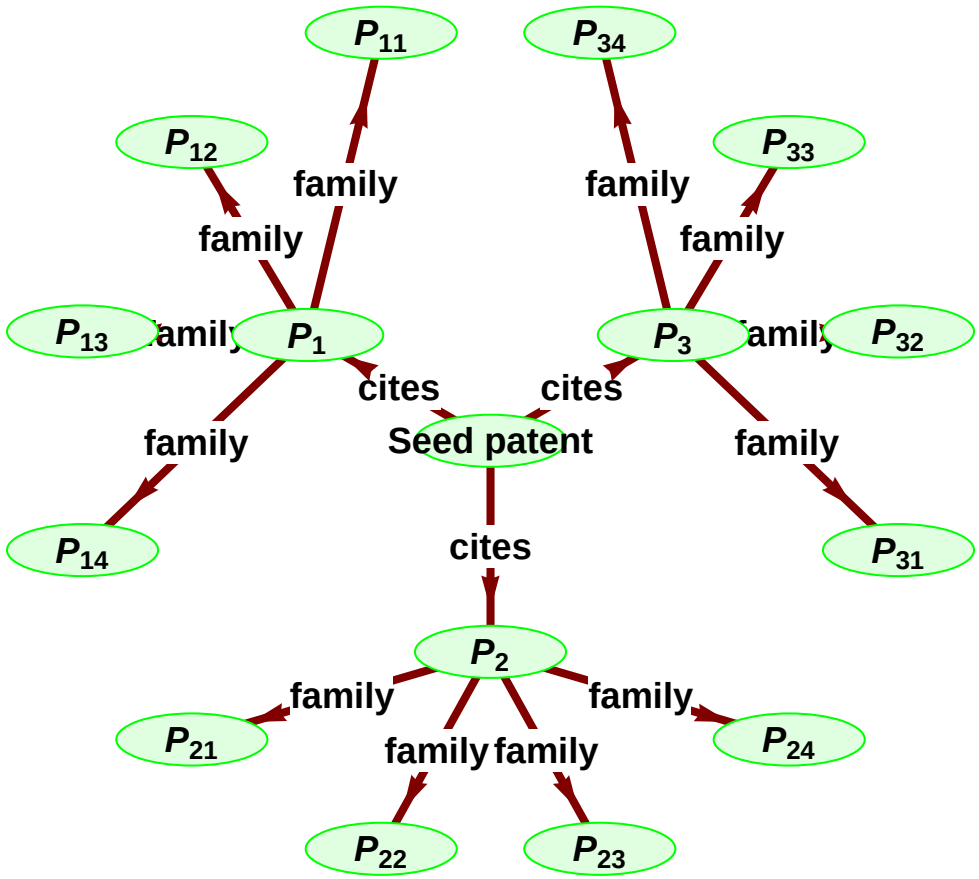
Relevance assessments

We used patents cited as prior art as relevance assessments.

Sources of citations:

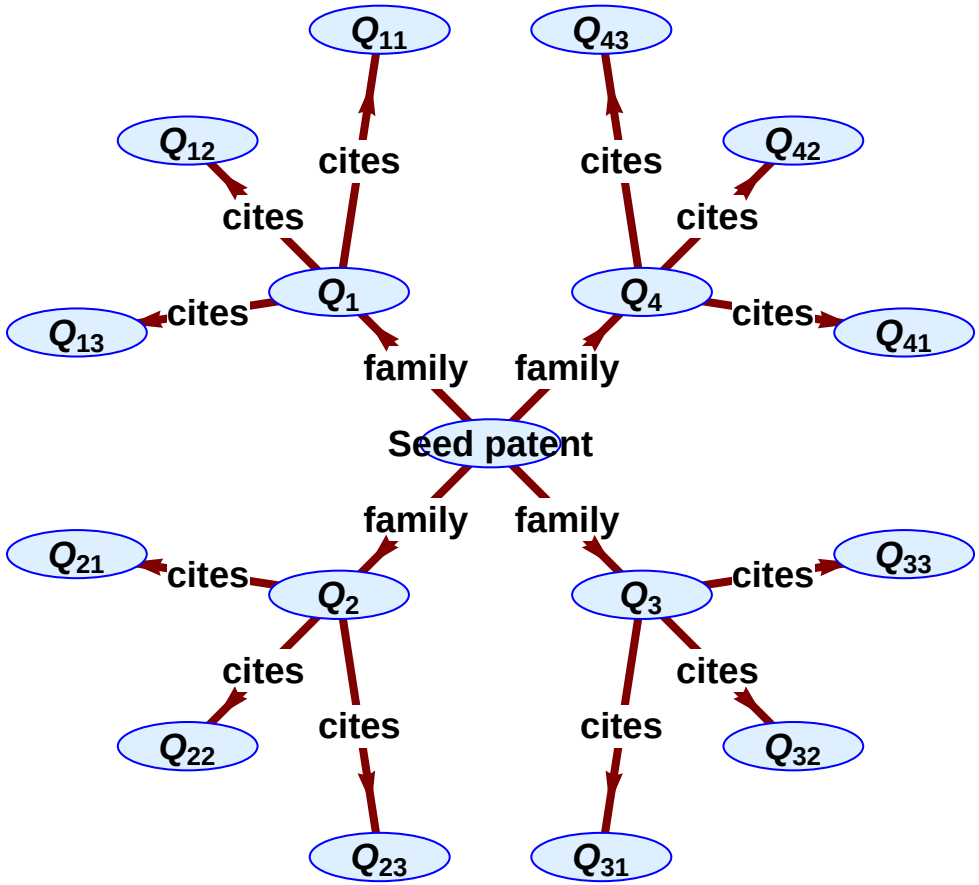
1. applicant's disclosure: the USPTO requires applicants to disclose all known relevant publications
2. patent office search report: each patent office will do a search for prior art to judge the novelty of a patent
3. opposition procedures: patents cited to prove that a granted patent is not novel

Extended citations as relevance assessments



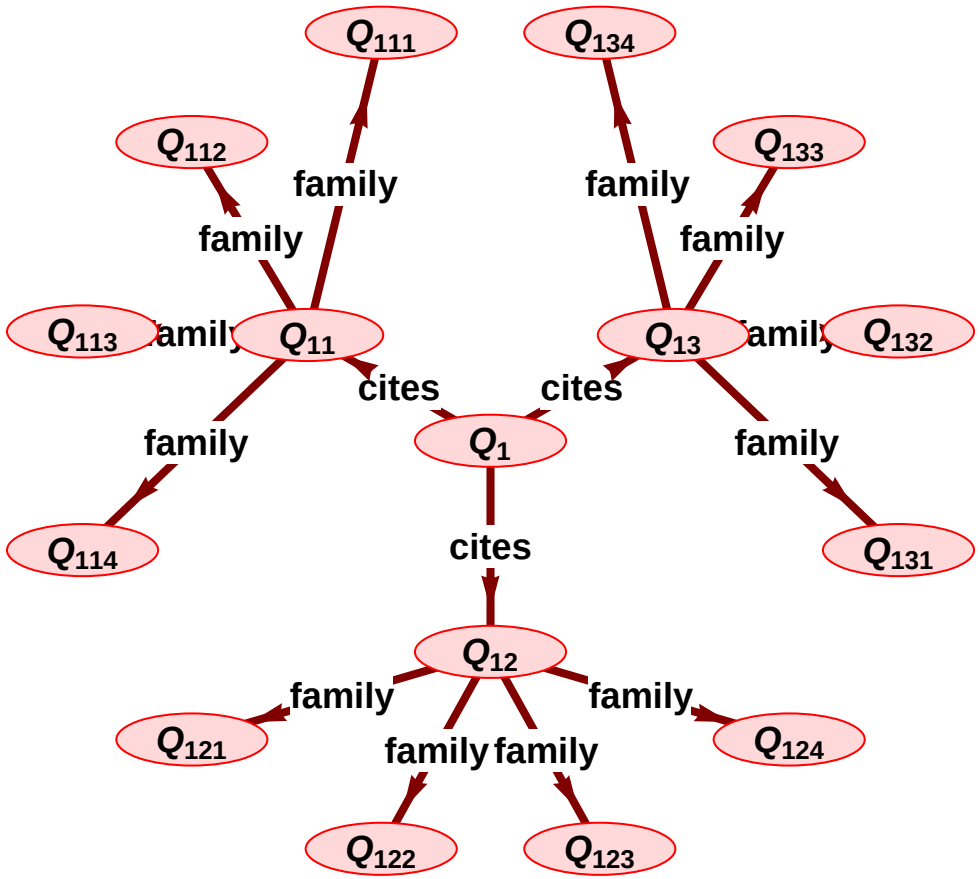
direct citations and their families

Extended citations as relevance assessments



direct citations of family members ...

Extended citations as relevance assessments



... and their families

Patent families

A patent family consists of patents granted by different patent authorities but related to the same invention.

Patent families

A patent family consists of patents granted by different patent authorities but related to the same invention.

simple family all family members share the same *priority number*

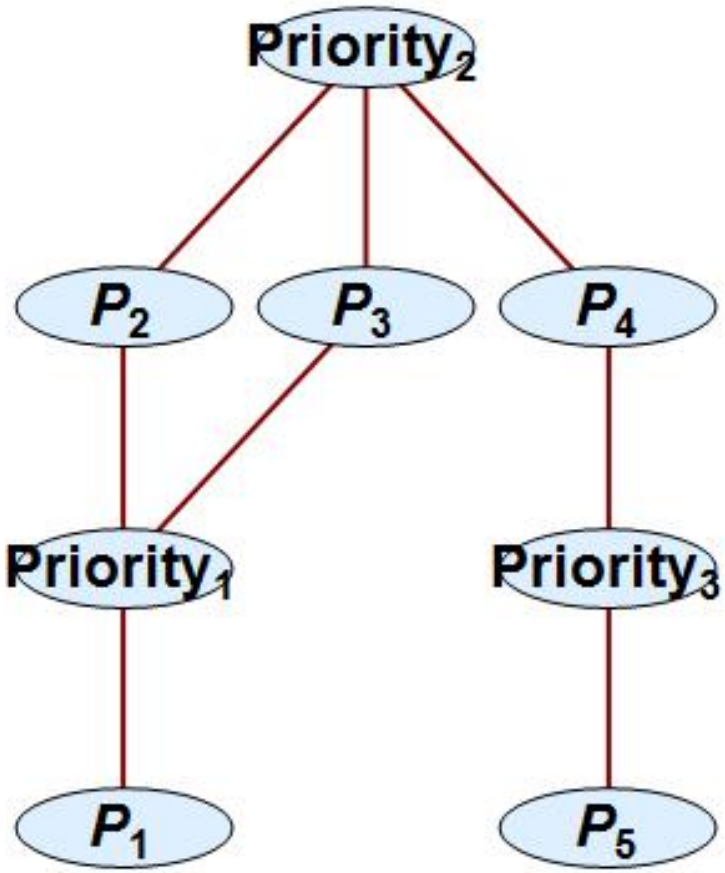
Patent families

A patent family consists of patents granted by different patent authorities but related to the same invention.

simple family all family members share the same *priority number*

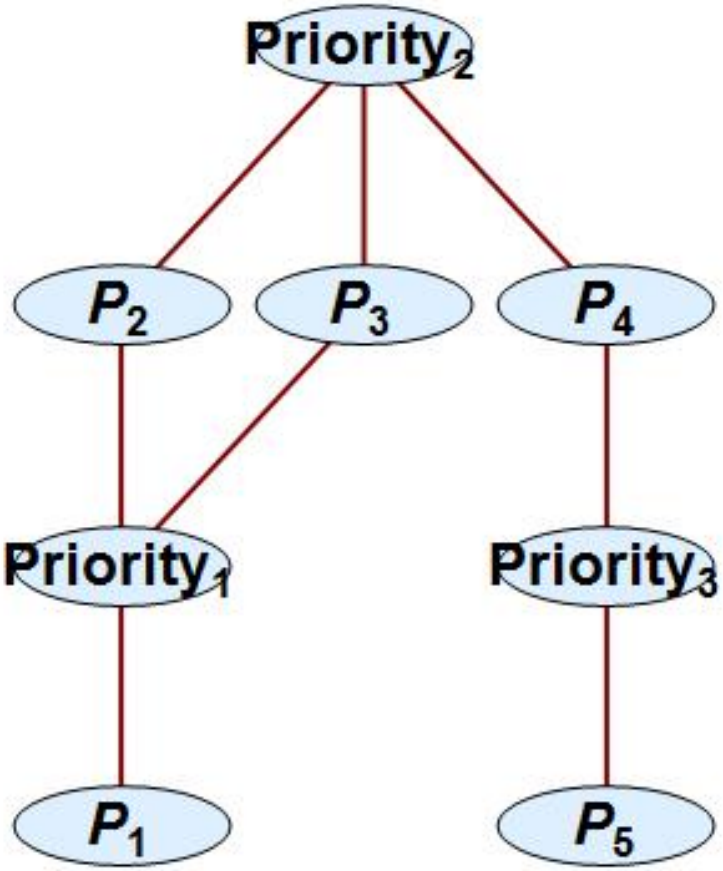
extended family there are several definitions, in the INPADOC database all documents which are directly or indirectly linked via a priority number belong to the same family

Patent families

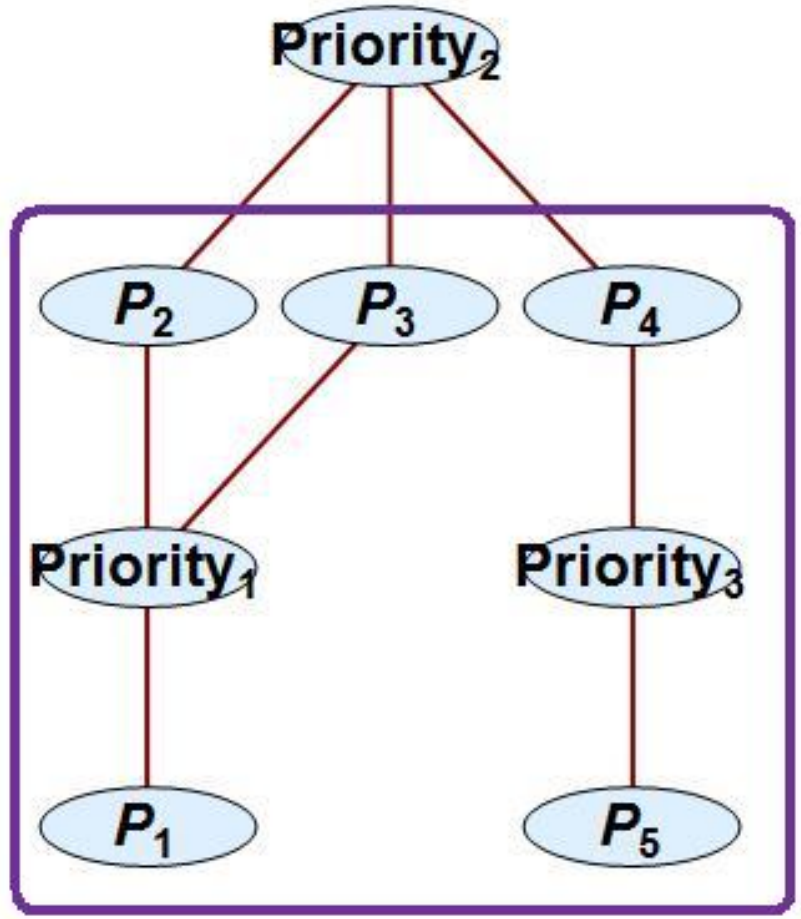


Patent documents are linked by priorities

Patent families

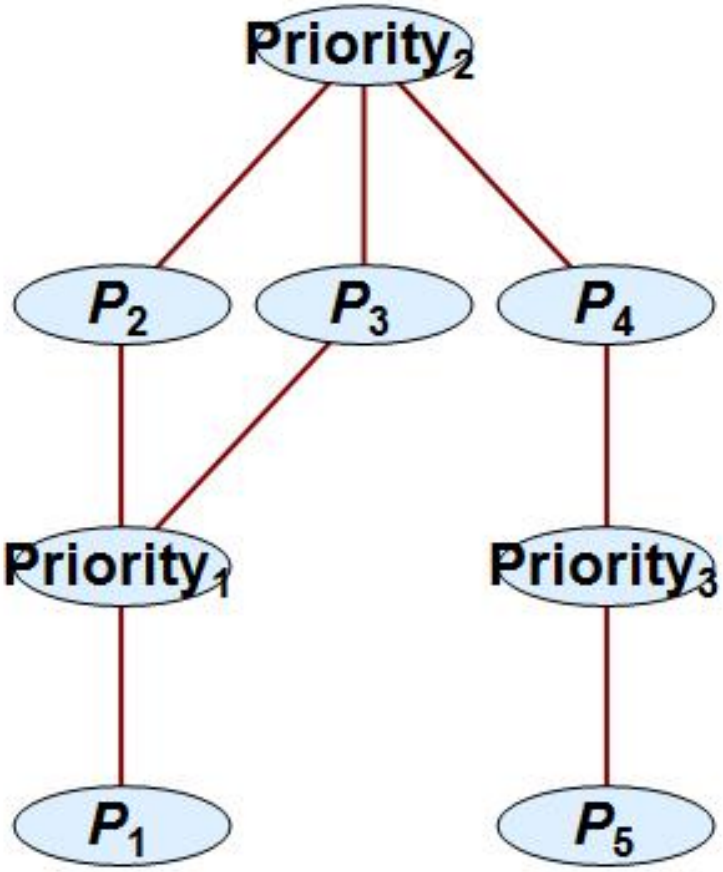


Patent documents are linked by priorities

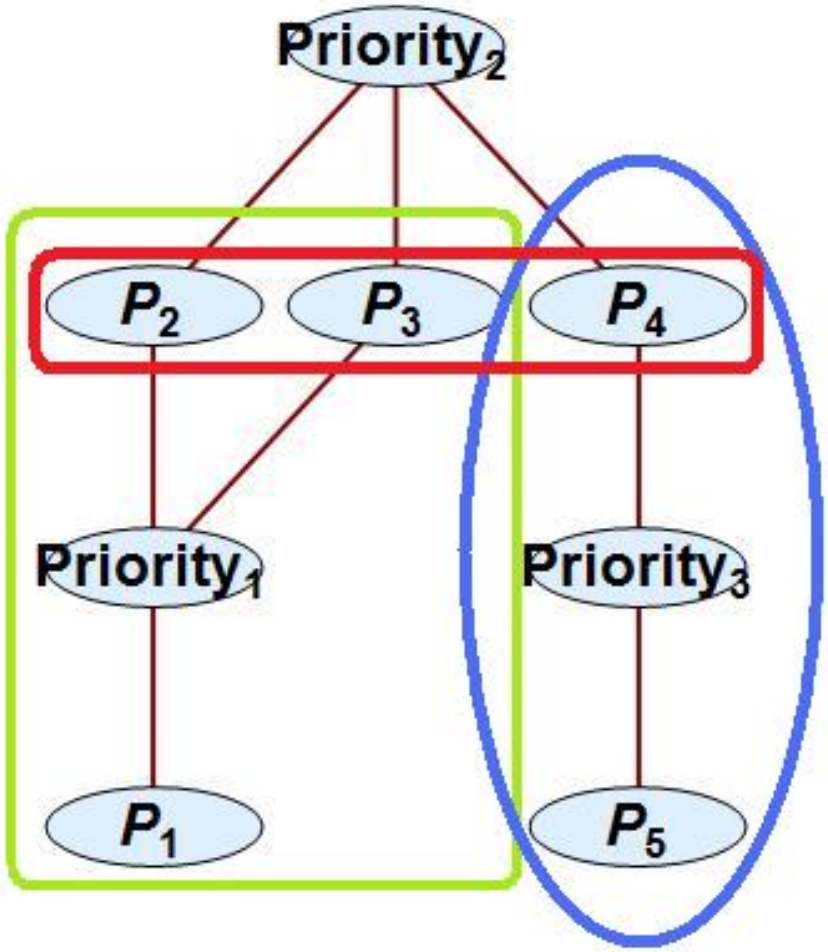


INPADOC family.

Patent families



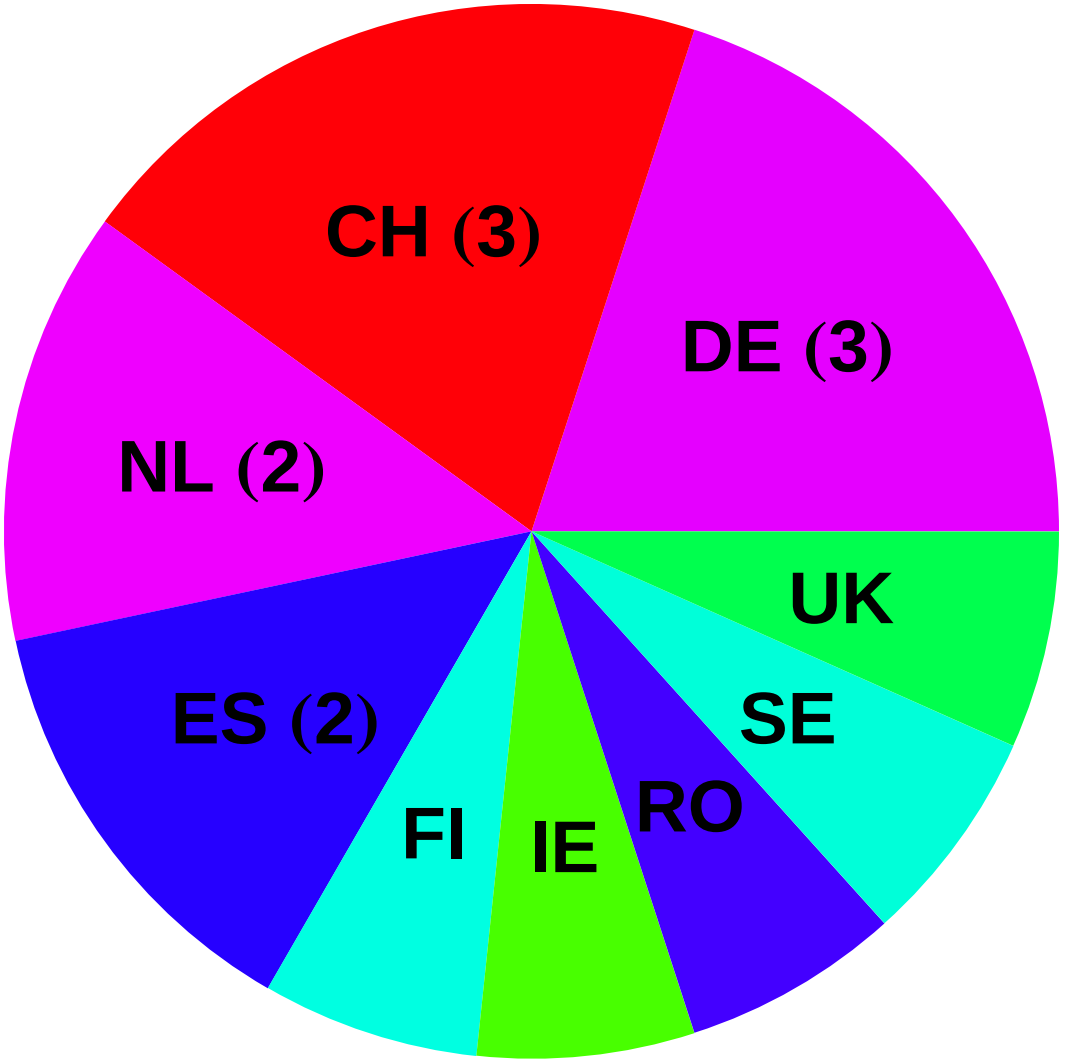
Patent documents are linked by priorities



CLEF–IP uses simple families.

Outline

Participants



- 15 participants
- 48 runs for the main task
- 10 runs for the language tasks

Participants

1. Tech. Univ. Darmstadt, Dept. of CS, Ubiquitous Knowledge Processing Lab (**DE**)
2. Univ. Neuchatel - Computer Science (**CH**)
3. Santiago de Compostela Univ. - Dept. Electronica y Computacion (**ES**)
4. University of Tampere - Info Studies (**FI**)
5. Interactive Media and Swedish Institute of Computer Science (**SE**)
6. Geneva Univ. - Centre Universitaire d'Informatique (**CH**)
7. Glasgow Univ. - IR Group Keith (**UK**)
8. Centrum Wiskunde & Informatica - Interactive Information Access (**NL**)

Participants

9 Geneva Univ. Hospitals - Service of Medical Informatics (**CH**)

10 Humboldt Univ. - Dept. of German Language and Linguistics (**DE**)

11 Dublin City Univ. - School of Computing (**IE**)

12 Radboud Univ. Nijmegen - Centre for Language Studies & Speech Technologies (**NL**)

13 Hildesheim Univ. - Information Systems & Machine Learning Lab (**DE**)

14 Technical Univ. Valencia - Natural Language Engineering (**ES**)

15 Al. I. Cuza University of Iasi - Natural Language Processing (**RO**)

Upload of experiments

A system based on Alfresco[2] together with a Docasu[3] web interface was developed.
Main features of this system are:

[2] http://www.alfresco.com/
[3] http://docasu.sourceforge.net/

Upload of experiments

A system based on Alfresco[2] together with a Docasu[3] web interface was developed.
Main features of this system are:

- user authentication

[2] http://www.alfresco.com/
[3] http://docasu.sourceforge.net/

Upload of experiments

A system based on Alfresco[2] together with a Docasu[3] web interface was developed.
Main features of this system are:

- user authentication
- run files format checks

[2] http://www.alfresco.com/
[3] http://docasu.sourceforge.net/

Upload of experiments

A system based on Alfresco[2] together with a Docasu[3] web interface was developed.
Main features of this system are:

- user authentication
- run files format checks
- revision control

[2] http://www.alfresco.com/
[3] http://docasu.sourceforge.net/

Who contributed

These are the people who contributed to the CLEF–IP track:

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee:

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee: Gianni Amati, Kalervo Järvelin, Noriko Kando, Mark Sanderson, Henk Thomas, Christa Womser-Hacker

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee: Gianni Amati, Kalervo Järvelin, Noriko Kando, Mark Sanderson, Henk Thomas, Christa Womser-Hacker
- Helmut Berger who invented the name CLEF–IP

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee: Gianni Amati, Kalervo Järvelin, Noriko Kando, Mark Sanderson, Henk Thomas, Christa Womser-Hacker
- Helmut Berger who invented the name CLEF–IP
- Florina Piroi and Veronika Zenz who walked the walk

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee: Gianni Amati, Kalervo Järvelin, Noriko Kando, Mark Sanderson, Henk Thomas, Christa Womser-Hacker
- Helmut Berger who invented the name CLEF–IP
- Florina Piroi and Veronika Zenz who walked the walk
- the patent experts who helped with advice and with assessment of results

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee: Gianni Amati, Kalervo Järvelin, Noriko Kando, Mark Sanderson, Henk Thomas, Christa Womser-Hacker
- Helmut Berger who invented the name CLEF–IP
- Florina Piroi and Veronika Zenz who walked the walk
- the patent experts who helped with advice and with assessment of results
- the Soire team

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee: Gianni Amati, Kalervo Järvelin, Noriko Kando, Mark Sanderson, Henk Thomas, Christa Womser-Hacker
- Helmut Berger who invented the name CLEF–IP
- Florina Piroi and Veronika Zenz who walked the walk
- the patent experts who helped with advice and with assessment of results
- the Soire team
- Evangelos Kanoulas and Emine Yilmaz for their advice on statistics

Who contributed

These are the people who contributed to the CLEF–IP track:

- the CLEF–IP steering committee: Gianni Amati, Kalervo Järvelin, Noriko Kando, Mark Sanderson, Henk Thomas, Christa Womser-Hacker
- Helmut Berger who invented the name CLEF–IP
- Florina Piroi and Veronika Zenz who walked the walk
- the patent experts who helped with advice and with assessment of results
- the Soire team
- Evangelos Kanoulas and Emine Yilmaz for their advice on statistics
- John Tait

Outline

Measures used for evaluation

We evaluated all runs according to standard IR measures

Measures used for evaluation

We evaluated all runs according to standard IR measures

- Precision, Precision@5, Precision@10, Precision@100

Measures used for evaluation

We evaluated all runs according to standard IR measures

- Precision, Precision@5, Precision@10, Precision@100
- Recall, Recall@5, Recall@10, Recall@100

Measures used for evaluation

We evaluated all runs according to standard IR measures

- Precision, Precision@5, Precision@10, Precision@100
- Recall, Recall@5, Recall@10, Recall@100
- MAP

Measures used for evaluation

We evaluated all runs according to standard IR measures

- Precision, Precision@5, Precision@10, Precision@100
- Recall, Recall@5, Recall@10, Recall@100
- MAP
- nDCG (with reduction factor given by a logarithm in base 10)

How to interpret the results

Some participants were disappointed by their poor evaluation results as compared to other tracks

How to interpret the results

MAP = 0.02 **?**

How to interpret the results

There are two main reasons why evaluation at CLEF–IP yields lower values than other tracks:

How to interpret the results

There are two main reasons why evaluation at CLEF–IP yields lower values than other tracks:

1. citations are incomplete sets of relevance assessments

How to interpret the results

There are two main reasons why evaluation at CLEF–IP yields lower values than other tracks:

1. citations are incomplete sets of relevance assessments
2. target data set is fragmentary, some patents are represented by one single document containing just title and bibliographic references (thus making it practically unfindable)

How to interpret the results

Still, one can sensibly use evaluation results for comparing runs assuming that

How to interpret the results

Still, one can sensibly use evaluation results for comparing runs assuming that

1. incompleteness of citations is distributed uniformly

How to interpret the results

Still, one can sensibly use evaluation results for comparing runs assuming that

1. incompleteness of citations is distributed uniformly
2. same assumption for unfindable documents in the collection

How to interpret the results

Still, one can sensibly use evaluation results for comparing runs assuming that

1. incompleteness of citations is distributed uniformly
2. same assumption for unfindable documents in the collection

Incompleteness of citations is difficult to check not having a large enough gold standard to refer to.

How to interpret the results

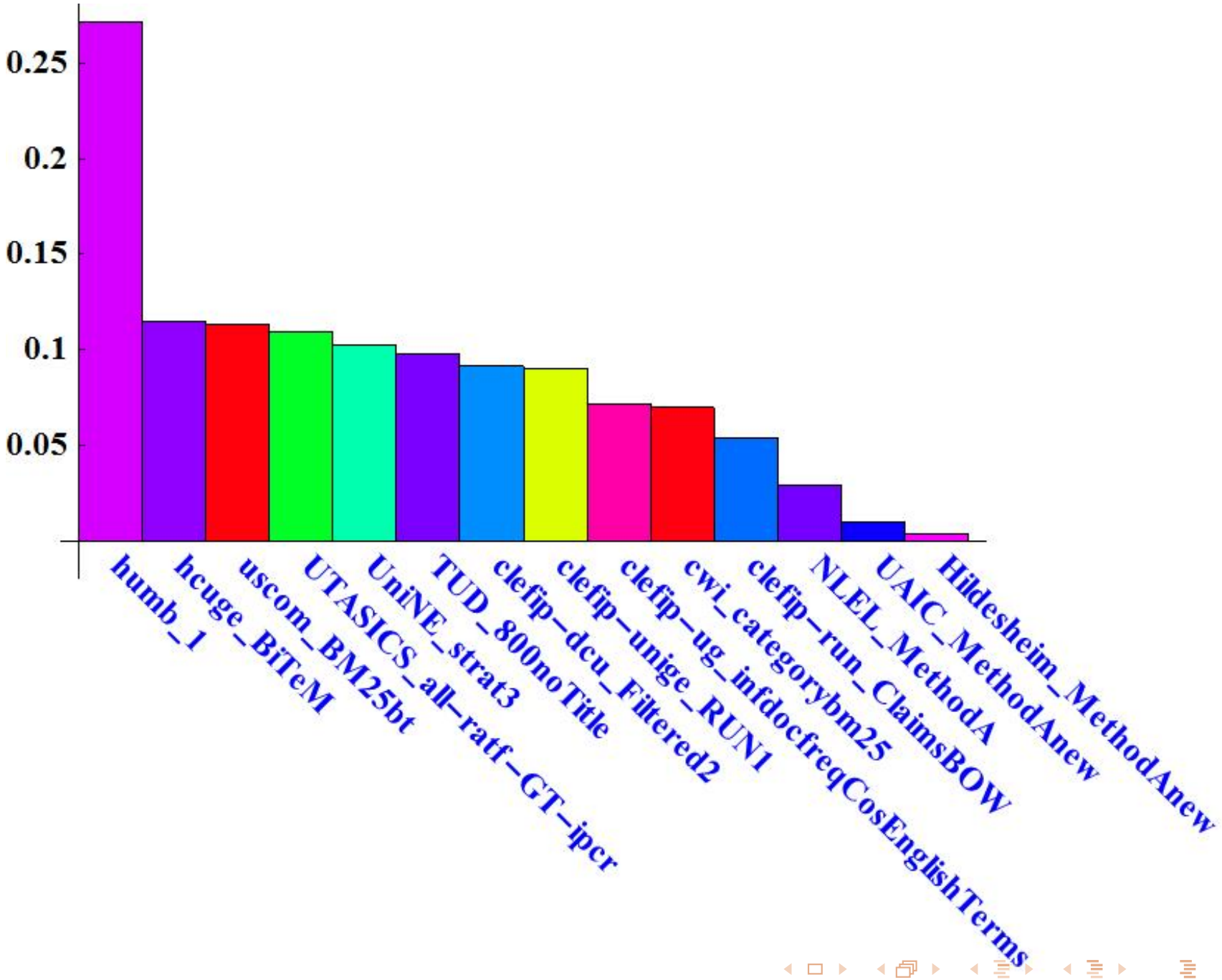
Still, one can sensibly use evaluation results for comparing runs assuming that

1. incompleteness of citations is distributed uniformly
2. same assumption for unfindable documents in the collection

Incompleteness of citations is difficult to check not having a large enough gold standard to refer to.

Second issue: we are thinking about re-evaluating all runs after removing unfindable patents from the collection.

MAP: best run per participant



MAP: best run per participant

Group-ID	Run-ID	MAP	R@100	P@100
humb	1	0.27	0.58	0.03
hcuge	BiTeM	0.11	0.40	0.02
uscom	BM25bt	0.11	0.36	0.02
UTASICS	all-ratf-ipcr	0.11	0.37	0.02
UniNE	strat3	0.10	0.34	0.02
TUD	800noTitle	0.11	0.42	0.02
clefip-dcu	Filtered2	0.09	0.35	0.02
clefip-unige	RUN3	0.09	0.30	0.02
clefip-ug	infdocfreqCosEnglishTerms	0.07	0.24	0.01
cwi	categorybm25	0.07	0.29	0.02
clefip-run	ClaimsBOW	0.05	0.22	0.01
NLEL	MethodA	0.03	0.12	0.01
UAIC	MethodAnew	0.01	0.03	0.00
Hildesheim	MethodAnew	0.00	0.02	0.00

Table: MAP, P@100, R@100 of best run/participant (S)

Manual assessments

We managed to have 12 topics assessed up to rank 20 for all runs.

Manual assessments

We managed to have 12 topics assessed up to rank 20 for all runs.

- 7 patent search professionals

Manual assessments

We managed to have 12 topics assessed up to rank 20 for all runs.

- 7 patent search professionals
- judged in average 264 documents per topics

Manual assessments

We managed to have 12 topics assessed up to rank 20 for all runs.

- 7 patent search professionals
- judged in average 264 documents per topics
- not surprisingly, rankings of systems obtained with this small collection do not agree with rankings obtained with large collection

Manual assessments

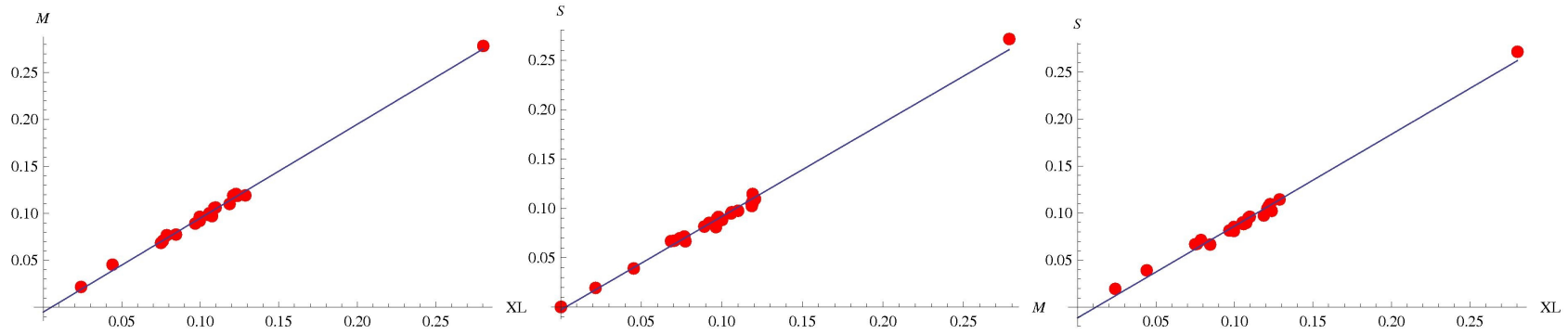
We managed to have 12 topics assessed up to rank 20 for all runs.

- 7 patent search professionals
- judged in average 264 documents per topics
- not surprisingly, rankings of systems obtained with this small collection do not agree with rankings obtained with large collection

Investigations on this smaller collection are ongoing.

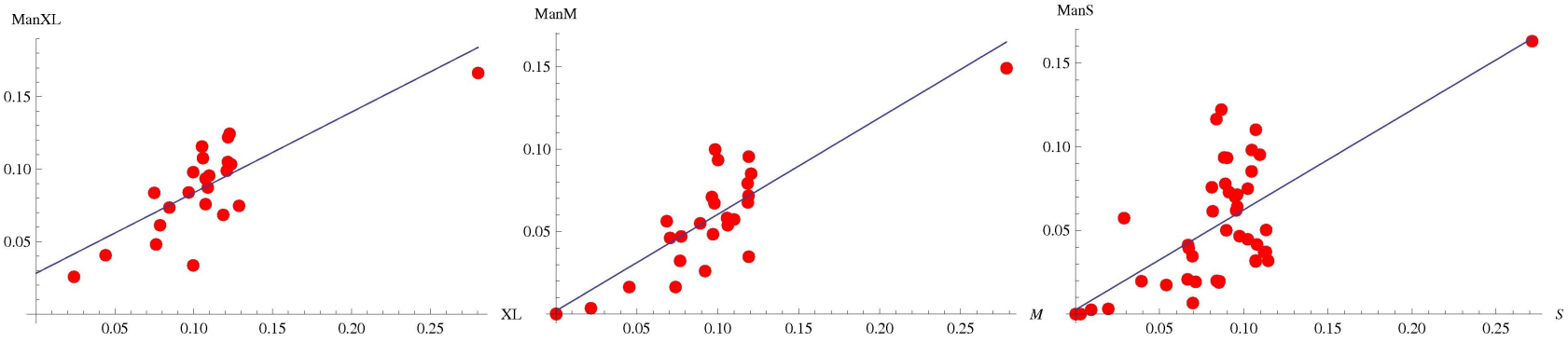
Correlation analysis

The rankings of runs obtained with the three sets of topics (**S**=500 ,**M**=1000, **XL**=10,000)are highly correlated (Kendall's $\tau > 0.9$) suggesting that the three collections are equivalent.



Correlation analysis

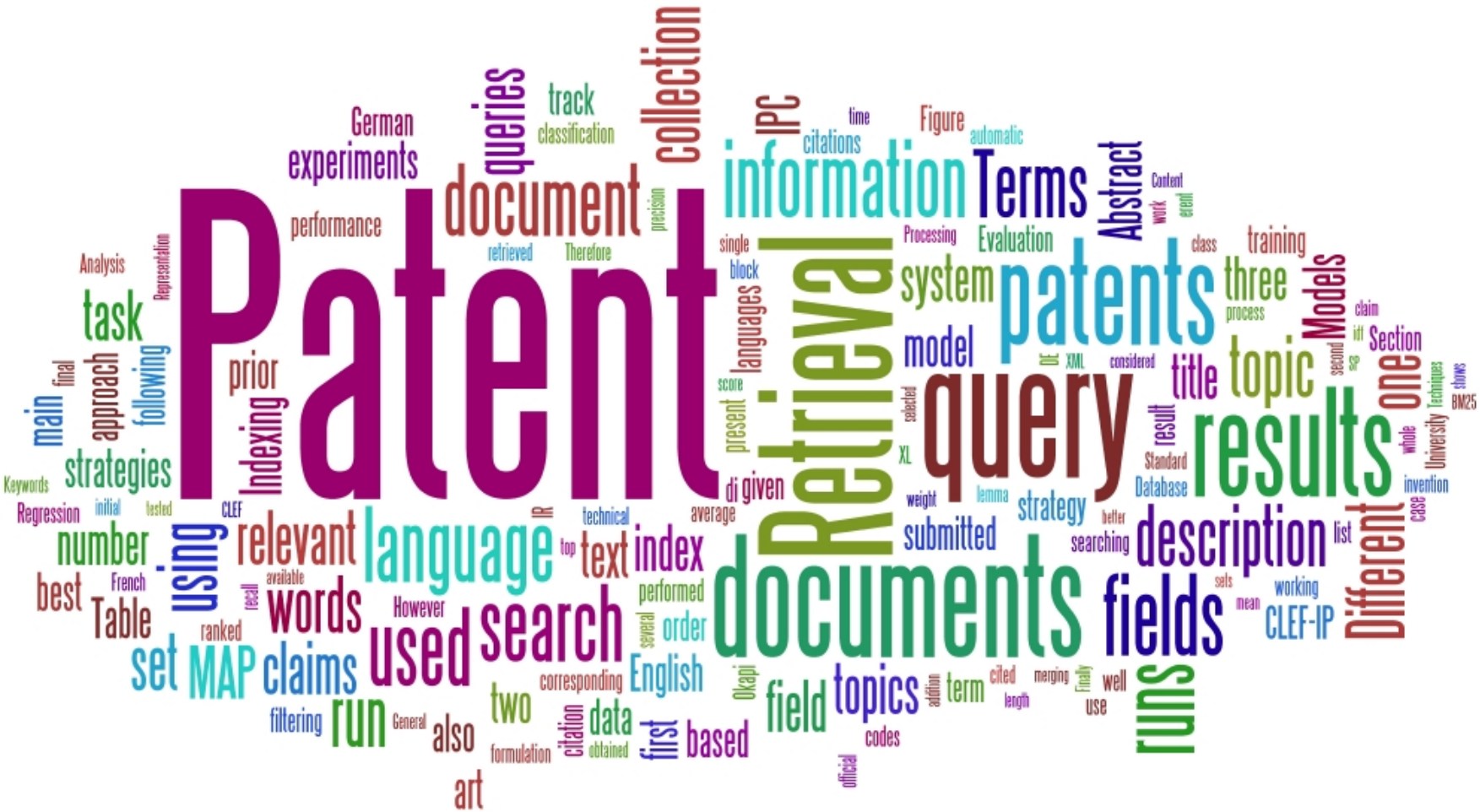
As expected, correlation drops when comparing the ranking obtained with the 12 manually assessed topics and the one obtained with the ≥ 500 topics sets.



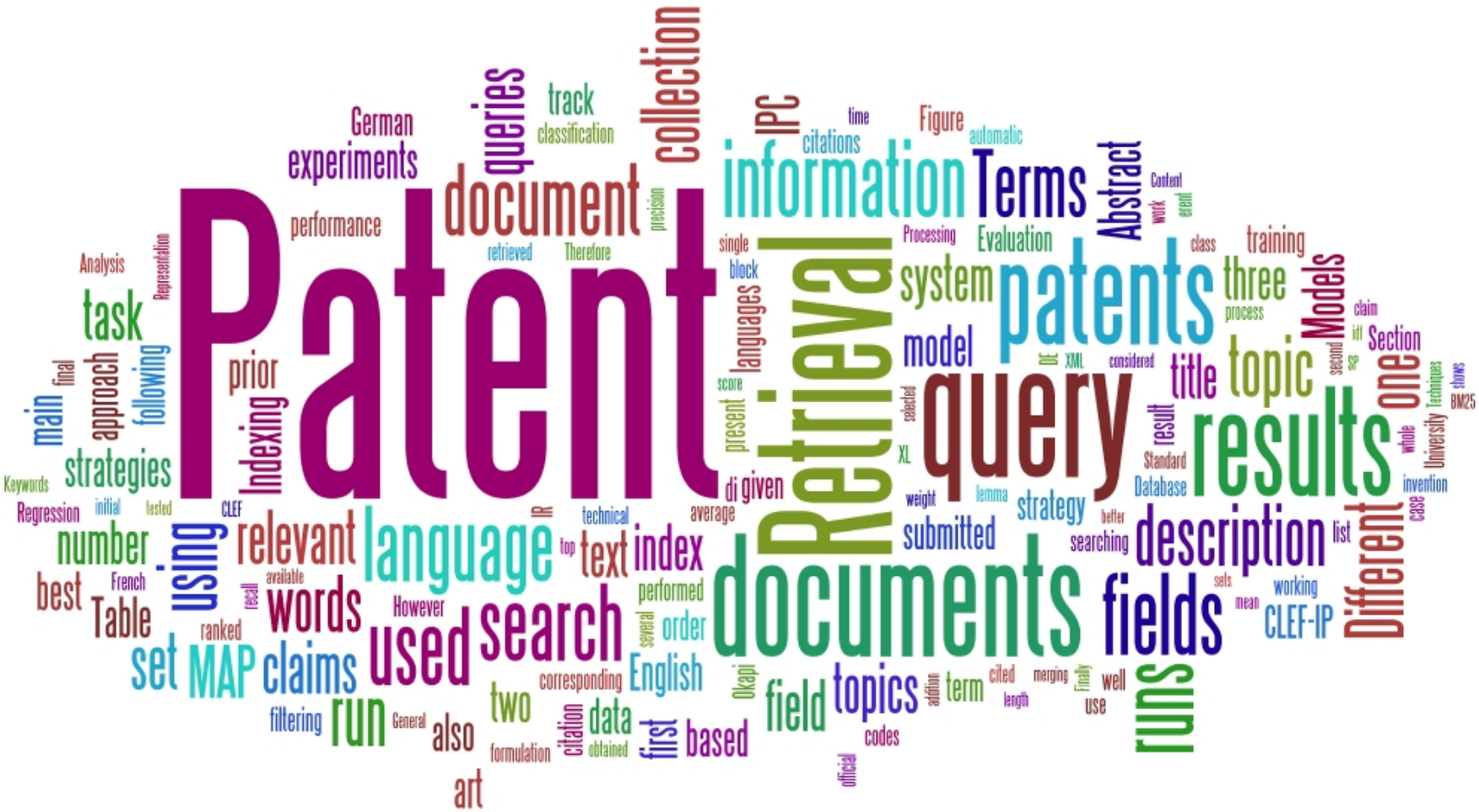
Working notes

I didn't have time to read the working notes ...

... so I collected all the notes and generated a Wordle



... so I collected all the notes and generated a Wordle



They're about patent retrieval.

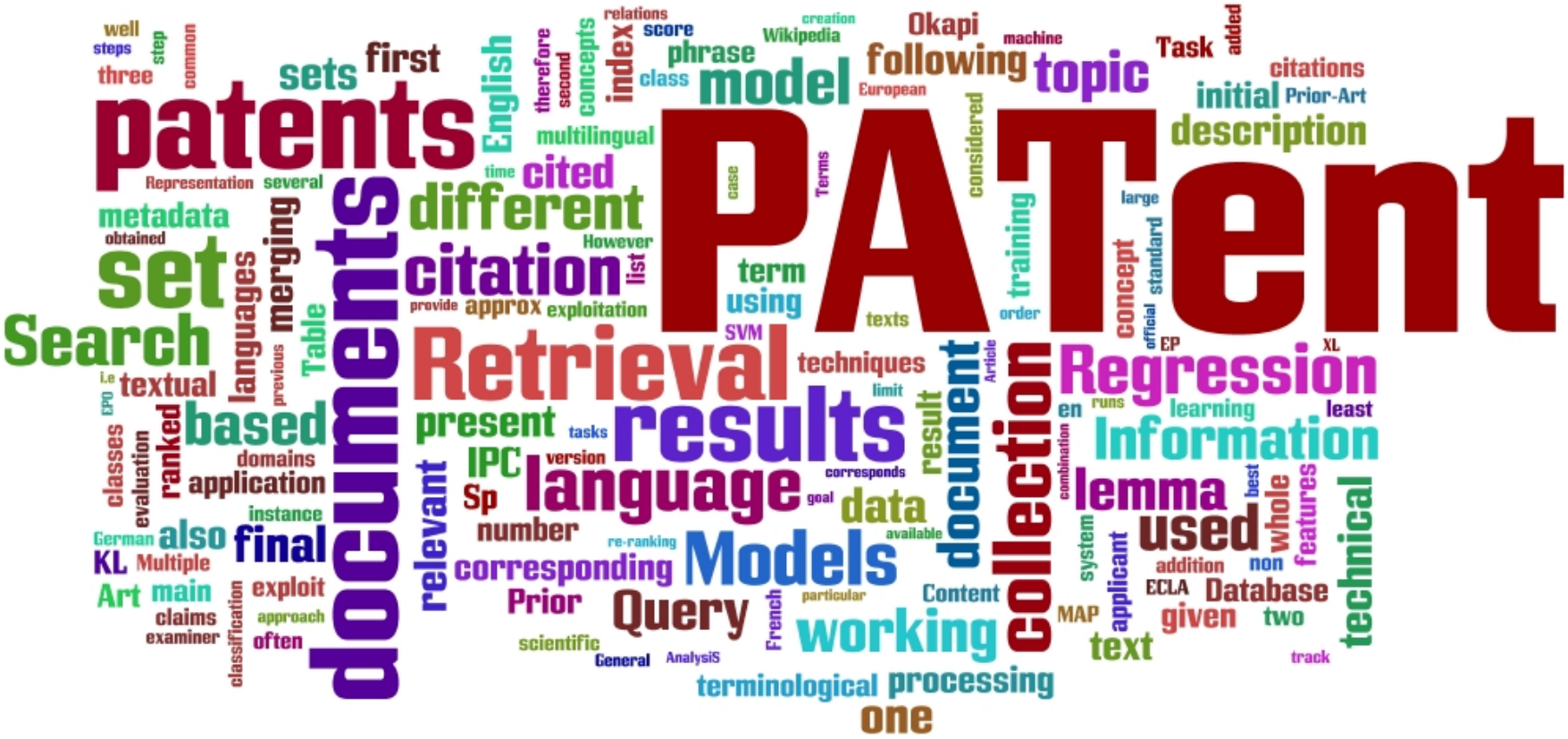
Refining the Wordle

I ran an Information Extraction algorithm in order to get a more meaningful picture

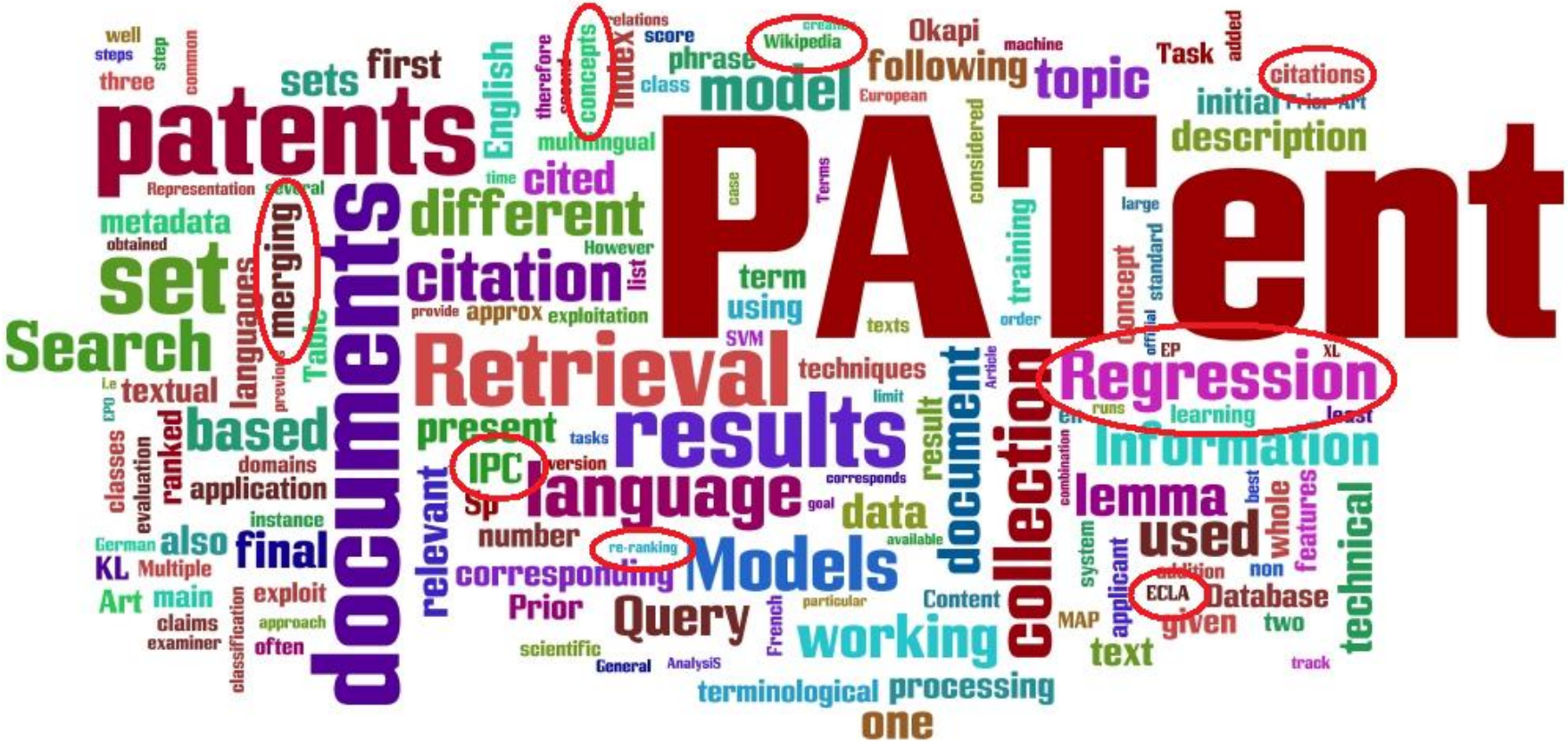
Refining the Wordle

Refining the Wordle

Humboldt's University working notes



Humboldt's University working notes



Humboldt's University working notes

Humboldt's University working notes

Humboldt's University working notes

Future plans

Some plans and ideas for future tracks:

Future plans

Some plans and ideas for future tracks:

- a layered evaluation model is needed in order to measure the impact of each single factor to retrieval effectiveness

Future plans

Some plans and ideas for future tracks:

- a layered evaluation model is needed in order to measure the impact of each single factor to retrieval effectiveness
- provide images (they are essential elements in chemical or mechanical patents, for instance)

Future plans

Some plans and ideas for future tracks:

- a layered evaluation model is needed in order to measure the impact of each single factor to retrieval effectiveness
- provide images (they are essential elements in chemical or mechanical patents, for instance)
- investigate query reformulations rather than one query-result set

Future plans

Some plans and ideas for future tracks:

- a layered evaluation model is needed in order to measure the impact of each single factor to retrieval effectiveness
- provide images (they are essential elements in chemical or mechanical patents, for instance)
- investigate query reformulations rather than one query-result set
- extend collection to include other languages

Future plans

Some plans and ideas for future tracks:

- a layered evaluation model is needed in order to measure the impact of each single factor to retrieval effectiveness
- provide images (they are essential elements in chemical or mechanical patents, for instance)
- investigate query reformulations rather than one query-result set
- extend collection to include other languages
- include an annotation task

Future plans

Some plans and ideas for future tracks:

- a layered evaluation model is needed in order to measure the impact of each single factor to retrieval effectiveness
- provide images (they are essential elements in chemical or mechanical patents, for instance)
- investigate query reformulations rather than one query-result set
- extend collection to include other languages
- include an annotation task
- include a categorization task

Epilogue

- we have created a large integrated test collection for experimentations in patent retrieval

Epilogue

- we have created a large integrated test collection for experimentations in patent retrieval
- the CLEF–IP track had a more than satisfactory participation rate for its first year

Epilogue

- we have created a large integrated test collection for experimentations in patent retrieval
- the CLEF–IP track had a more than satisfactory participation rate for its first year
- the right combination of techniques and the exploitation of patent-specific know-how yields best results

Thank you for your attention.