Information Retrieval Facility

CLEF-IP: Information Retrieval in Intellectual Property Domain

Florina Piroi, John Tait

September 20, 2010

CLEF-IP at CLEF

Information Retrieval Facility

- Launched in 2009
- Aim: investigate IR methods in patent retrieval
- Focus: Cross–language retrieval for European languages

The CLEF-IP 2010 Collection

Information Retrieval Facility

- Target data: 2.6 million documents, representing 1.9 million patents (EPO, Marec);
- Prior Art Candidate Search Task: 2000 topics
- Classification Task: 2000 topics
- Relevance Assessments: based on patent citations and existing patent classification

The CLEF-IP 2010 Collection

Information Retrieval Facility

- Target data: 2.6 million documents, representing 1.9 million patents (EPO, Marec);
- Prior Art Candidate Search Task: 2000 topics
- Classification Task: 2000 topics
- Relevance Assessments: based on patent citations and existing patent classification

Target data: standardized XML format for patent data (Marec Scheme), with multilingual content in English, French and German.

Patents, what about them?

Information Retrieval Facility

A patent is a set of exclusive legal rights for the use and exploitation of an invention in exchange for its public disclosure.

- legal documents
- granted by a governing authority
- rights limited in time

Main Steps towards Granted Patents

Information Retrieval Facility

Application

Provide written specification (*Application Document*) which contains

- background of the invention (narrative)
- description of the invention (narrative)
- set of claims (scope of protection) (attornish)

Application Date or *Filing Date*

Main Steps towards Granted Patents

Information Retrieval Facility

Application

Examination

Application analyzed by professionals for:

- novelty
- inventiveness
- practicability

Novelty Search $\equiv$ Prior Art Search

- time consuming
- needs expert knowledge

Main Steps towards Granted Patents

Information Retrieval Facility

Application

Examination

Prior Art Search $\rightsquigarrow$ **Search Report**

Documents relevant to the application: **Patent Citations**

Main Steps towards Granted Patents

Information Retrieval Facility

Application

Application Document

Examination

Prior Art Search $\rightsquigarrow$ *Search Report*

Negotiation $\rightsquigarrow$ Patent Grant

A *Granted Patent Document* is published, exclusive rights are granted.
Claims may change !
EPO: translation of the claims in English, German, French must be provided.

After the Grant

Information Retrieval Facility

Application

Application Document

Examination

Prior Art Search $\rightsquigarrow$ *Search Report*

Negotiation $\rightsquigarrow$ Patent Grant

Granted Patent Document

Opposition procedure

A prior art search for the *Patent Document* is done. *Invalidity Search* $\rightsquigarrow$ *Additional Search Report*

Patent Organization Tools

Information Retrieval Facility

Patent Families

- Same invention filed at more than one patent office
- *Priority Date* - first Application Date
- simple family, Inpadoc, Derwent

Patent Classification Systems

- 'Sorts' patents according to the technical area they belong to.
- Basis for quick technical investigation of a field
- IPC, ECLA, F-Terms, US Classification System

The CLEF-IP 2010 Collection

Information Retrieval Facility

- Target data: 2.6 million patent documents, representing 1.9 million patents (EPO, Marec);
- Prior Art Candidate Search Task, Novelty Search: Application Documents as topics
- Classification Task, IPC: Application Documents as topics
- Relevance Assessments: based on *extended* patent citations and existing IPC patent classification
- CLS Target data: set of IPC subclasses (post January 1st 2006)

The CLEF-IP 2010 Collection

Information Retrieval Facility

- Target data: 2.6 million patent documents, representing 1.9 million patents (EPO, Marec);
- Prior Art Candidate Search Task, Novelty Search: Application Documents as topics
- Classification Task, IPC: Application Documents as topics
- Relevance Assessments: based on *extended* patent citations and existing IPC patent classification
- CLS Target data: set of IPC subclasses (post January 1st 2006)

Previous Work

- NTCIR workshop series (2001) – Japanese, Chinese, English
- NIST, TREC-CHEM (2009) – Chemistry, English

Hello?

Information Retrieval Facility

Participants and Runs

Information Retrieval Facility

ID	Institution		CLS	PAC
bitem	BiTeM, Service of Medical Informatics, Geneva University Hospitals	CH	7	2
dcu	Dublin City Univ. - School of Computing	IE		3
hild	Hildesheim Univ. - Information Science	DE		4
humb	Humboldt Univ. - Dept. of German Language and Linguistics	DE	1	1
insa	LCI, Institut National des Sciences Appliquèes de Lyon	FR	5	
jve	Industrial Property Documentation Dept., JSI Jouve	FR	3	
run	Information Foraging Lab, Radboud Univ. Nijmegen	NL	2	2
spq	Spinque	NL	1	1
ssft	Simple Shift	CH	8	
uaic	Al. I. Cuza University of Iași - NLP	RO		1
ui	IR Group, Universitas Indonesia	ID		3
uned	UNED - E.T.S.I. Informatica, Dpto. Lenguajes y Sistemas Informaticos	ES		8
Total	**12**		**27**	**25**

Measures

PAC Task

- Precision, Precision@5, Precision@10, Precision@50, Precision@100
- Recall, Recall@5, Recall@10, Recall@50, Recall@100
- MAP
- nDCG
- PRES[a]

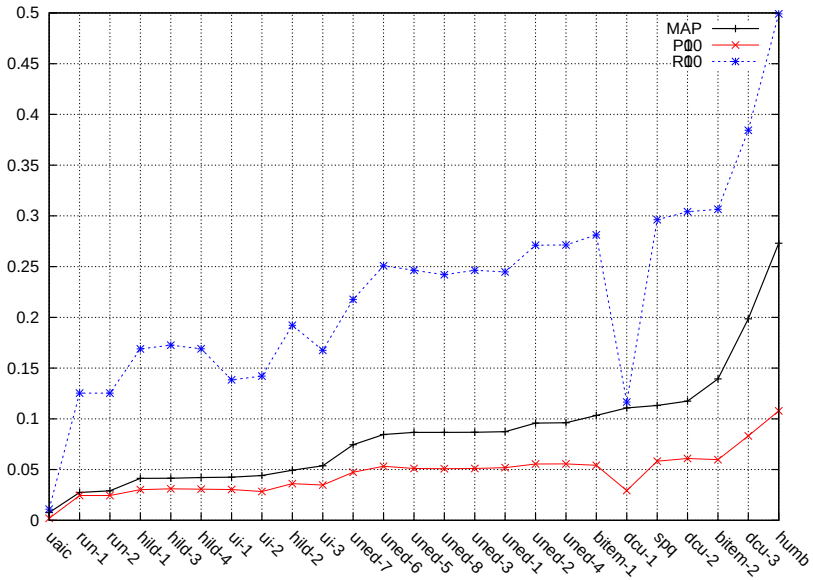
[a] Magdy W. and G. J. F. Jones. PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. SIGIR 2010

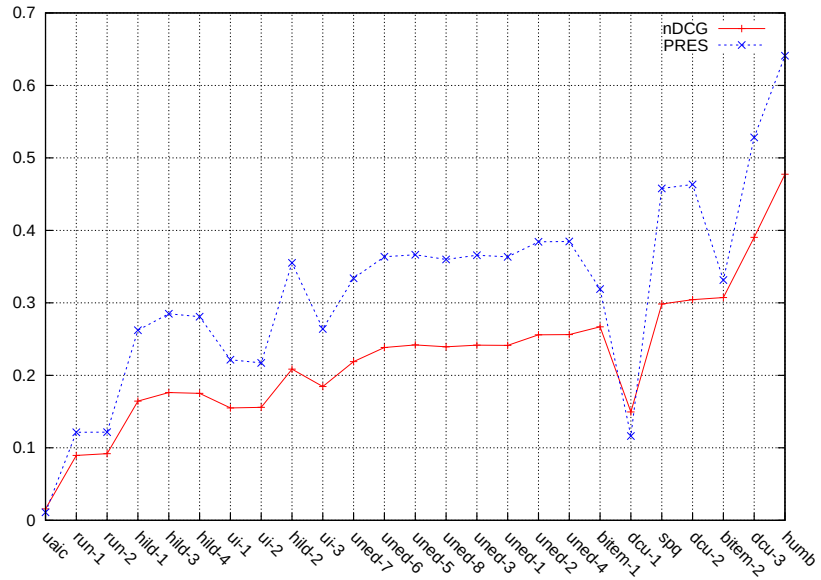
CLS Task

- Precision@5, Precision@10, Precision@25, Precision@50
- Recall@5, Recall@25, Recall@50,
- MAP, F_1 at 5, 25 and 50.

Prior Art Candidates Search Task

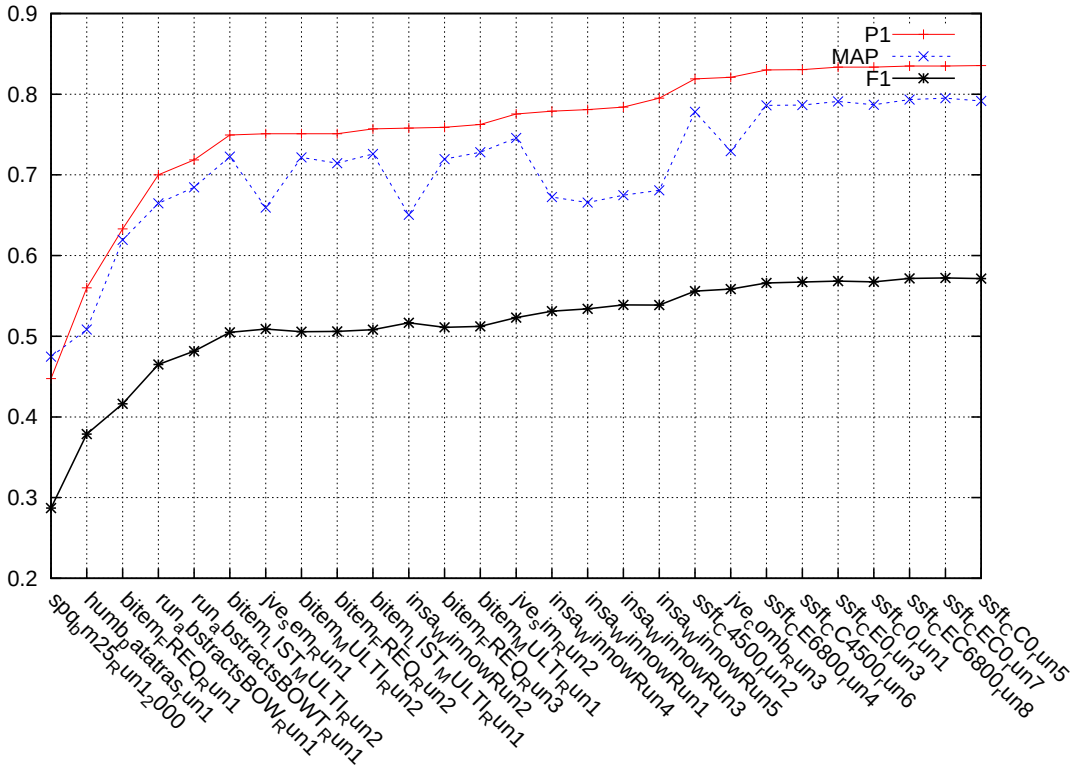
Information Retrieval Facility





Classification Task

Information Retrieval Facility



Future Plans

Information Retrieval Facility

- other evaluation models
- manual assessments
- include images
- query–reformulation
- add more European languages
- refine the tasks
- better defined multi–lingual tasks
- better involvement of IP professionals
- ...

Information Retrieval Facility

Acknowledgments

Information Retrieval Facility

Advisory Board

- Kalervo JÄRVELIN (University of Tamperem FI)
- Noriko KANDO (National Institute of Informatics, JP)
- Javier POSE RODRÍGUEZ (European Patent Office, DE)
- Giovanna RODA (Information Retrieval Specialist, AT)
- Mark SANDERSON (University of Sheffield, UK)
- Anthony TRIPPE (3LP Advisors, US)
- Christa WOMSER-HACKER (Hildesheim University, DE)

Acknowledgments

Information Retrieval Facility

Advisory Board

CLEF-IP Participants

CLEF Conference and Lab Organizers

IRF Colleagues https://www.ir-facility.org/about/people/staff

Thank You

Information Retrieval Facility

Wednesday, September 22nd.
CLEF-IP Workshop