

TU  
WIEN

!

FAKULTÄT  
FÜR INFORMATIK  
  
Faculty of Informatics

Domain Specific IR

Lecture 3 of 5: Patent IR

Mihai Lupu

lupu@ifs.tuwien.ac.at

Russian Summer School on Information Retrieval

August 6-11, 2012

Yaroslavl, Russian Federation

Monolingual Text

Just because it's English – it doesn't have to be English

What is claimed is:

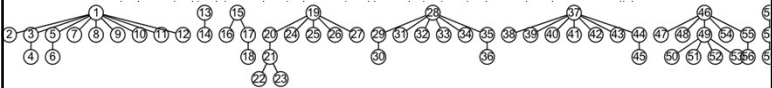
1. A method for scrolling through portions of a data set, said method comprising: receiving a number of units associated with a rotational user input; determining an acceleration factor pertaining to the rotational user input; modifying the number of units by the acceleration factor; determining a next portion of the data set based on the modified number of units; and presenting the next portion of the data set.

2. A method as recited in claim 1, wherein the data set pertains to a list of items, and the portions of the data set include one or more of the items.

3. A method as recited in claim 1, wherein the data set pertains to a media file, and the portions of the data set pertain to one or more sections of the media file.

4. A method as recited in claim 3, wherein the media file is an audio file.

5. A method as recited in claim 1, wherein the rotational user input is provided via a rotational input device.



10. A method as recited in claim 1, wherein said determining of the next data portion comprises: converting the modified number of units into the next portion based on a predetermined value.

11. A method as recited in claim 1, wherein said determining of the next data portion comprises: dividing the modified number of units by a chunking value.

12. A method as recited in claim 1, wherein said determining of the next data portion comprises: adding a prior remainder value to the modified number of units; and converting the modified number of units into the next portion.

Monolingual Text

Just because it's English – it doesn't have to be English

What is claimed is:

1. A method for scrolling through portions of a data set, said method comprising: receiving a number of units associated with a rotational user input; determining an acceleration factor pertaining to the rotational user input; modifying the number of units by the acceleration factor; determining a next portion of the data set based on the modified number of units; and presenting the next portion of the data set.

2. A method as recited in claim 1, wherein the data set pertains to a list of items, and the portions of the data set include one or more of the items.

3. A method as recited in claim 1, wherein the data set pertains to a media file, and the portions of the data set pertain to one or more sections of the media file.

4. A method as recited in claim 3, wherein the media file is an audio file.

5. A method as recited in claim 1, wherein the rotational user input is provided via a rotational input device.

6. A method as recited in claim 5, wherein the rotational input device is a circular touch pad or a rotary dial.

7. A method as recited in claim 1, wherein the acceleration factor is dependent upon a rate of speed for the rotational user input.

8. A method as recited in claim 1, wherein the acceleration factor provides a range of acceleration.

9. A method as recited in claim 1, wherein the acceleration factor can successively increase to provided successively greater levels of acceleration.

10. A method as recited in claim 1, wherein said determining of the next data portion comprises: converting the modified number of units into the next portion based on a predetermined value.

11. A method as recited in claim 1, wherein said determining of the next data portion comprises: dividing the modified number of units by a chunking value.

12. A method as recited in claim 1, wherein said determining of the next data portion comprises: adding a prior remainder value to the modified number of units; and converting the modified number of units into the next portion.

Monolingual Text

Just because it's English – it doesn't have to be English

What is claimed is:

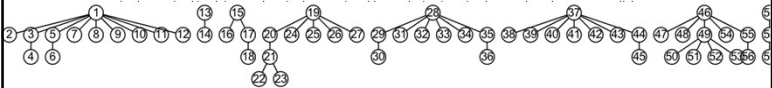
1. A method for scrolling through portions of a data set, said method comprising: receiving a number of units associated with a rotational user input; determining an acceleration factor pertaining to the rotational user input; modifying the number of units by the acceleration factor; determining a next portion of the data set based on the modified number of units; and presenting the next portion of the data set.

2. A method as recited in claim 1, wherein the data set pertains to a list of items, and the portions of the data set include one or more of the items.

3. A method as recited in claim 1, wherein the data set pertains to a media file, and the portions of the data set pertain to one or more sections of the media file.

4. A method as recited in claim 3, wherein the media file is an audio file.

5. A method as recited in claim 1, wherein the rotational user input is provided via a rotational input device.



10. A method as recited in claim 1, wherein said determining of the next data portion comprises: converting the modified number of units into the next portion based on a predetermined value.

11. A method as recited in claim 1, wherein said determining of the next data portion comprises: dividing the modified number of units by a chunking value.

12. A method as recited in claim 1, wherein said determining of the next data portion comprises: adding a prior remainder value to the modified number of units; and converting the modified number of units into the next portion.

NOT FOR REDISTRIBUTION

1

### Monolingual Text

- Just because it's English – it doesn't have to be English
- What is claimed is:
  1. A method for scrolling through portions of a data set, said method comprising: receiving a number of units associated with a rotational user input; determining an acceleration factor pertaining to the rotational user input; modifying the number of units by the acceleration factor; determining a next portion of the data set based on the modified number of units; and presenting the next portion of the data set.
  2. A method as recited in claim 1, wherein the data set pertains to a list of items, and the portions of the data set include one or more of the items.
  3. A method as recited in claim 1, wherein the data set pertains to a media file, and the portions of the data set pertain to one or more sections of the media file.
  4. A method as recited in claim 3, wherein the media file is an audio file.
  5. A method as recited in claim 1, wherein the rotational user input is provided via a rotational input device.

This was from the Application (WO).  
The EP-B (granted patent) has only 35 claims.

- 10. A method as recited in claim 2, wherein said determining of the next data portion comprises: dividing the modified number of units by a chunking value.
- 11. A method as recited in claim 2, wherein said determining of the next data portion comprises: adding a prior remainder value to the modified number of units; and converting the modified number of units into the next portion.

### Monolingual Text

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

7

### Monolingual Text

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

6

### Monolingual text

- It is no longer plain English
  - Do the assumptions about the distribution of words still hold? → does TF/IDF still hold?
  - Not necessarily [Sarasua:2000]
    - Drop the tf
    - Calculate the idf only at class level
    - Introduce pip (position in phrase) weight

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

8

Monolingual Text

- Compare different weighting/scoring techniques
- models that perform well on news corpora (BM25, log(tf).idf.id), perform well on the patent corpora too, relative to the other models

Table 2.1. Retrieval models analysed by Iwayama et al.

Model	weight
hits	$b_{q,t} \times b_{d,t}$
baseline	$f_{q,t} \times b_{d,t}$
tf	$f_{q,t} \times \frac{f_{d,t}}{df_d}$
idf	$f_{q,t} \times idf_t$
tf.idf	$f_{q,t} \times idf_t \times \frac{f_{d,t}}{df_d}$
log(tf)	$(1 + \log(f_{q,t})) \times \frac{1 + \log(f_{d,t})}{1 + \log(ave f_d)}$
log(tf).idf	$(1 + \log(f_{q,t})) \times idf_t \times \frac{1 + \log(f_{d,t})}{1 + \log(ave f_d)}$
log(tf).idf.idf	$(1 + \log(f_{q,t})) \times idf_t \times \frac{1 + \log(f_{d,t})}{1 + \log(ave f_d)} \times \frac{1}{ave df_d + S \times (df_d - ave df_d)}$
BM25	$f_{q,t} \times \log \left( \frac{N - n_t + 0.5}{n_t + 0.5} \right) \times \frac{\frac{f_{d,t}}{b_{d,t} \times idf_t}}{K \times \left( (1 - b) + b \times \frac{f_{d,t}}{b_{d,t} \times idf_t} \right) + f_{d,t}}$

[Iwayama et al. : 2003]

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

9

Document Length

- Patent documents are longer than news corpora.
- Why?
- Normally, one of two causes:
  - Unitary topic, but verbose
  - Multiple topics
- Patent document = 1 invention = 1 topic
- Not always
- “divisional” application, USPTO “continuation” & “continuation in part”

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

11

Monolingual Text

- Follow up study [Fujita:2005]
  - BM25-variant vs. language modelling
  - Focus on the effects of document length
  - Result:
    - Retrieval improved when the model penalizes long documents
    - BM25: set  $b$  to higher values (0.9 – 1.0 suggested for the patent domain, compared to 0.3 – 0.4 for news corpora)

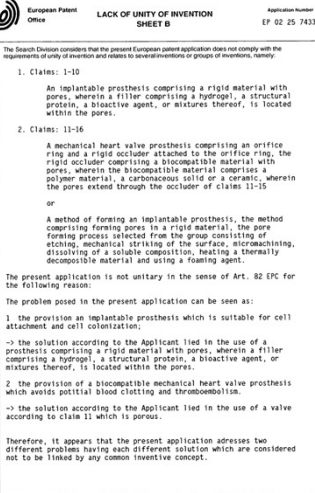
RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

10

Document Length

- Patent documents are longer than news corpora.
- Why?
- Normally, one of two causes:
  - Unitary topic, but verbose
  - Multiple topics
- Patent document = 1 invention = 1 topic
- Not always
- “divisional” application, USPTO “continuation” & “continuation in part”



RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

12

Monolingual Text

- The lack-of-unity = problem search prior art for an application
- Try automatic topic detection
- [Ganguly:2011] uses TextTiling

Run Name	Parameters			Evaluation metric		
	Segmented Fusion method			PRES	MAP	Recall
WHOLE	No	N/A		0.4413	0.0899	0.5310
SEG_COMBSUM	Yes	COMBSUM		0.1545	0.0308	0.1759
SEG_RR	Yes	Round-robin		<b>0.4949</b>	<b>0.0947</b>	<b>0.5982</b>

- and Pseudo Relevance Feedback (PDF)

Run Name	Parameters			Evaluation metric			
	Segmented	PRF	R	T	PRES	MAP	Recall
WHOLE	No	No	-	-	0.4413	0.0899	0.5310
WHOLE_PRF	Yes	Yes	10	10	0.4415	0.0889	0.5333
SEG	Yes	No	-	-	0.4949	0.0947	0.5982
SEG_PRF	Yes	Yes	10	10	<b>0.5033</b>	<b>0.1025</b>	<b>0.6166</b>

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

Monolingual text

- Extracting queries from patents
  - Often requests for information=full patent or claim
  - [Xue:2009] propose a method to extract keywords from patents for prior art
  - Based on a learning to rank approach
  - 3 types of features
    - Retrieval-score:num, field, weight, NP
    - Low-level: variants of tfidf
    - Category: from classification codes

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

15

Monolingual Text

- [Mahdabi:2011] improves upon it using Language Modelling, and different query lengths (25 .. 150)

PQM(desc)	25	50	75	100	125	150
MAP	0.08	0.09	0.09	<b>0.10</b>	0.10	0.09
Recall	0.56	0.57	0.59	<b>0.59</b>	0.58	0.57

Using the Description field

PQM(clm)	25	50	75	100	125	150
MAP	0.04	0.05	0.06	<b>0.07</b>	0.07	0.07
Recall	0.48	0.50	0.52	<b>0.54</b>	0.53	0.52

CBQM(desc)	25	50	75	100	125	150
MAP	0.08	0.09	0.10	<b>0.11</b>	0.10	0.09
Recall	0.58	0.59	0.60	<b>0.60</b>	0.59	0.59

LLQM(desc)	25	50	75	100	125	150
MAP	0.08	0.08	0.11	<b>0.12</b>	0.12	0.11
Recall	0.59	0.62	0.62	<b>0.63</b>	0.61	0.60

PQM(clm)	25	50	75	100	125	150
MAP	0.05	0.06	0.06	<b>0.07</b>	0.06	0.06
Recall	0.49	0.52	0.53	<b>0.56</b>	0.54	0.52

LLQM(clm)	25	50	75	100	125	150
MAP	0.06	0.08	0.10	<b>0.10</b>	0.09	0.09
Recall	0.51	0.53	0.56	<b>0.57</b>	0.56	0.55

Using the Claims field

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

Monolingual text

- Extracting queries from patents

ALGORITHM: Transforming Patent to Query  
INPUT: Patent, Num, Field, Weight, NP  
OUTPUT: Query  
PROCESS: Rank words in Field according to their tfidf scores and then select Num top ranked words as the query words. Assign Weight to each query word to get Query<sub>w</sub>. Repeat the above steps for non-phrases to get Query<sub>np</sub>. If NP is true, set Query as the combination of Query<sub>w</sub> and Query<sub>np</sub>; otherwise set Query as Query<sub>w</sub>.

Figure 1: General algorithm for transforming the query patent to an effective search query.

Table 3: Influence of Weight on retrieval performance.

Field	MAP			Recall@100		
	local	tfidf	tf	local	tfidf	tf
ttl	0.042	0.039	0.043	0.143	0.129	0.144
drwd	0.044	0.048	0.047	0.133	0.144	0.143
detd	0.053	0.057	0.066*	0.167	0.171	0.169*
pcldm	0.059	0.062	0.055	0.179	0.183	0.167
clms	0.066	0.066	0.064	0.194	0.195	0.187
abst	0.066	0.070	0.074*	0.194	0.195	0.207*
all	0.067	0.068	0.078*	0.193	0.198	0.215*
bsum	0.078	0.082	0.094**	0.223	0.231	0.252**

Table 4: Influence of NP on retrieval performance

Field	MAP			Recall@100		
	w	w+p	w+p+np	w	w+p	w+p+np
ttl	0.043	0.042	0.144	0.137		
drwd	0.047	0.048	0.143	0.145		
detd	0.066	0.066	0.189	0.187		
pcldm	0.055	0.056	0.167	0.168		
clms	0.064	0.066*	0.187	0.191		
abst	0.074	0.074	0.207	0.208		
all	0.078	0.080*	0.215	0.219*		
bsum	0.094	0.096*	0.252	0.256		

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

16

# Monolingual Text

- Latent Semantic Indexing
  - Some commercial systems use it
    - <http://www.freepatentsonline.com>
    - "Latent semantic analysis uses sophisticated statistical analysis of language to search on concepts, not just words, to help you find those documents - even if they don't contain any of the words you used in your search"
    - [Riley:2008]
  - Minimal improvements found in experiments
    - [Moldovan:2005]

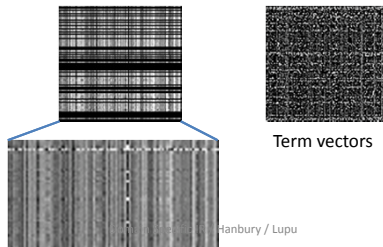
RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

17

# Random Indexing

- Initial experiments using the Semantic Vectors package
  - Unsatisfactory results for document similarity
  - Noticeably good results for term similarity



Document vectors

Term vectors

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

19

# Monolingual Text

- Latent Semantic Indexing
  - Some commercial systems use it
    - <http://www.freepatentsonline.com>
    - "Latent semantic analysis uses sophisticated statistical analysis of language to search on concepts, not just words, to help you find those documents - even if they don't contain any of the words you used in your search"
    - [Riley:2008]
  - Minimal improvements found in experiments
    - [Moldovan:2005]

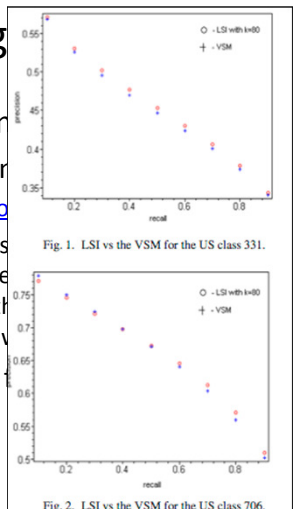


Fig. 1. LSI vs the VSM for the US class 331.

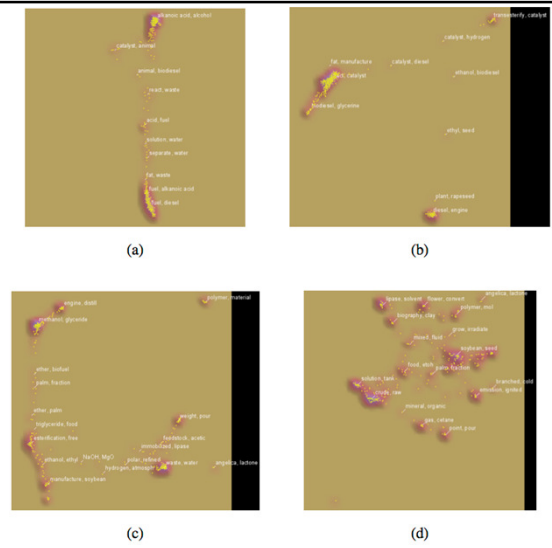
Fig. 2. LSI vs the VSM for the US class 706.

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

18

- Stop words



(a)

(b)

(c)

(d)

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

20

Monolingual Text

- Stop words
  - Manually created by domain experts
  - Automatically created
    - In general
      - Based on text statistics
        - » E.g. in Terrier
      - Evolutionary
        - » Genetic algorithms [Sinka:2003]
    - For patents in particular
      - [Kern:2011] – although view from the opposite side of finding discriminating words

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

21

Monolingual Text

- Out-of-vocabulary issue
- How much is the patent corpus covered by the CELEX lexical database?

	Patent data	COBUILD corpus
Tokens	96%	92%
Types	55%	(?)

- Most frequent out-of-vocabulary (other than numbers: *indicio, U-shaped, cross-section, cross-sectional, flip-flop, L-shaped, spaced-apart, thyristor, cup-shaped, and V-shaped.*
- patent claims do not use many words that are not covered by a lexicon of general English

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

23

Monolingual Text

- More than Bag-of-words – NLP on patents
- Most work on the claims section
- [Verberne:2010] – 67292 Claims vs BNC
  - Average claims length: 54 (median: 22) words
  - Sentences up to 3684 and 5089 words occur
  - High type/token ratio
    - Use of many different words
  - High Hapax ratio
    - (the proportion of terms that occur only once)
    - Lack of repetition

• % of Marec sentences    • % of BNC sentences

% of sentences in corpus

sentence length

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

22

Monolingual Text

- Use the SPECIALIST lexicon to identify multi-word terms
  - 200k 2-word terms, 30k 3-word terms and 10k 4-or-more-word terms
- Coverage:
  - <2% for 2-word terms
  - <1% for 3-word terms
- Most frequent: *carbon atoms, alkyl group, hydrogen atom, amino acid, molecular weight, combustion engine, control device, nucleic acid, semiconductor device and storage means*
- Introduction of ad-hoc multi-word terms is common and general practice

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

24

Monolingual Text

- Syntactic Structure
  - 1 sentence
  - Claims are Noun Phrases instead of Phrases

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

25

Monolingual Text

- Does NLP help in retrieval?
- Ambiguous results so far (as in other domains)

Run	Recall	Precision	MAP	P@5
EN_BM25_Terms_allFields	0.3298	0.0125	0.0414	0.0914
EM_BM25_Phrases_allFields	0.3605	0.0116	0.0422	0.0938
EM_BM25_Phrases(6)_title	0.4954	0.0118	0.0500	0.0844
Other CLEF-IP 2010 run using simple terms	0.57		0.1216	

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

27

Monolingual Text

- Syntactic Structure

Claims

- A component of an implantable valved prosthesis, the component comprising a rigid material with pores, a filler comprising a hydrogel or a structural protein, or mixtures thereof, located within the pores, and a surface cellular layer attached to a surface of the prosthesis, wherein said rigid porous material with the filler and the surface cellular layer presents a smooth surface for fluid flow.
- The prosthesis component of claim 1 in the form of an occluder.
- The prosthesis component of claims 1 and 2 wherein the rigid material extends through the rigid material.
- The prosthesis component of any preceding claim wherein the rigid material fills the pores.
- The prosthesis component of any preceding claim wherein the rigid material partly fills the pores.

head of Phrases

	AEGIR	Connexor CFG
precision	0.45	0.71
recall	0.50	0.71
F1-score	0.47	0.71
Inter-annotator agreement	0.83	0.83

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

26

Monolingual text

- Extracting queries from patents
  - Small parenthesis on NP use

Field	MAP		Recall100	
	w	w+p	w	w+p
ttl	0.043	0.042	0.144	0.137
drwd	0.047	0.048	0.143	0.145
dctd	0.066	0.066	0.189	0.187
pclaim	0.055	0.056	0.167	0.168
claim	0.064	0.066*	0.187	0.191
abst	0.074	0.074	0.207	0.208
all	0.078	0.080*	0.215	0.219*
best	0.094	0.096*	0.252	0.256

Tokens & NP & Entities	Tokens	Informative Noun Phrases	Non-informative Noun Phrases
0.4355	0.44	copper strip test methoxypropynyl group biodegradable collagen self-adhesive CODAL tape tyrosine kinase inhibitor	1 2 3 1 2 m 4 R=H 1200 W 13.56 MHz RF power about 1800 mg/kg A)1>[M M]/(4 [M M] [M M]) such difficulties

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

28



## Monolingual Text

- Perhaps we over-complicate things...
  - There exist basic patterns in claims
    - [Shinmori:2003] and [Sheremetyeva:2003] use keywords to identify relations (e.g. relations: *PROCEDURE, COMPONENT, ELABORATION, FEATURE, PRECONDITION, COMPOSE*)
    - Use them to split up the claims to help the [Stanford] parser.
      - [Parapatics:2009]
- |              | Heap Size | orig/split | successful |
|--------------|-----------|------------|------------|
| Training Set | 1000MB    | orig       | 80.4%      |
|              |           | split      | 99.5%      |
|              | 3000MB    | split      | 53.6%      |

	Heap Size	orig/split	successful parses
Training Set	1000MB	orig	80.4%
		split	<b>99.5%</b>
	500MB	orig	53.6%
		split	<b>96.2%</b>
Test Set	1000MB	orig	83.5%
		split	<b>98.0%</b>
	500MB	orig	52.4%
		split	<b>92.2%</b>

RuSSIR 2012, August 6-11

Domain Specific IR /

/ Hanbury / Lupu

## Monolingual Text

- **Information Extraction**
  - Because higher precision/recall is needed
  - Because of specific information needs
    - “mixtures with a melting temperature between 10C and 12C”
  - A lot of work done in the context of GATE @ Sheffield
  - [Cunningham:2011] The SAMIE components listed in runtime order (items in **bold** were developed specifically for SAMIE, other components were customised as needed)

1] The SAMIE components listed in runtime order (items in **bold** were developed specifically for SAMIE, other components were customised as needed)

Processing resource	Description
Cleanup	Remove annotations from previous application runs
Import Relevant Markup	Makes relevant markup from the original document available to the rest of the pipeline
<b>Roman Numerals</b>	Annotates Roman numerals which are used for detecting references
<b>Numbers in Words</b>	Recognises numbers written as words and converts them to actual values
Tokeniser	Pattern matcher for detection of words and other lexical items
Sentence splitter	Regular expression-based detection of sentence boundaries
POS tagger	Addition of part of speech (grammatical categories) to tokens
Gazetteer (case sensitive)	Lookup of known domain terms
Gazetteer (case insensitive)	Lookup of known domain terms, with case insensitive matching
<b>Numbers</b>	Find and annotate all remaining numbers
<b>References Transducers</b>	Find and annotate all the references within the documents <sup>30</sup>
<b>Measurement Tagger</b>	Find and annotate all the measurements within the documents

RuSSIR 2012, August 6-11

References Transducer Find and

I will annotate all the ref

## Monolingual Text

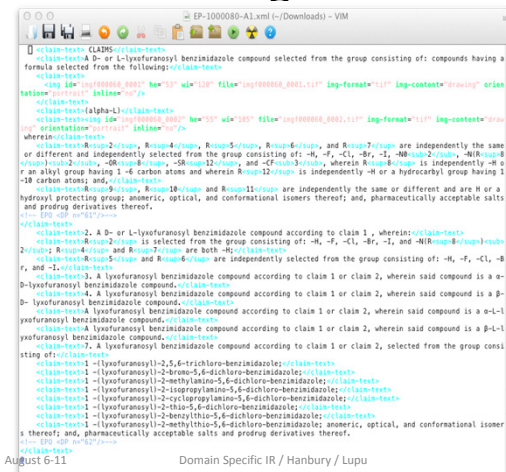
- Chemistry search
  - Particularly important due to commercial interest
  - Huge amount of manual indexing
    - E.g. Chemical Abstracts Service
  - [Emmerich:2009] studies the different results obtained by ‘first level’ and ‘second level’ patent sources
    - New documents found in every source

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

31

## Monolingual Text



RuSSIR 2012, August 6-11

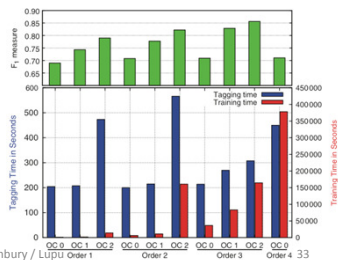
Domain Specific IR / Hanbury / Lupu

32



## Monolingual Text

- IUPAC names are popular
- Conditional Random Fields (CRFs) are popular to recognize them (according to BioCreative)
- [Klinger:2008] obtains a score up to 85% in terms of F1 measure
- [Grego:2009] compares CRF with dictionary approaches  
dictionary does better on partial matches – can be used as anchors



RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

## Multilinguality

- Cross-lingual search (querytranslation)
  - (fire AND protection) AND (building OR structure) AND NOT sprinkler
- Each keyword translated independently
  - But make use of tips in the query
    - (building OR structure) → you know which synset you need to look at
  - Not all keywords need to be translated
    - Pn:1234567 OR inventor:brown
  - Impossible to handle wild-cards

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

35

## Multilinguality



- Document translation
- Advantage of the domain:
  - Large amounts of comparable multilingual data
- Disadvantage: the language
  - Needs experts to verify translations
- Extensive use of translation memories
  - A multi-level dictionary (paragraph, phrase, sub-phrase)
- Use of English as Pivot is relatively common
- NTCIR-8 : showed for the first time that an SMT system can do better than a RBMT system for Japanese

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

34

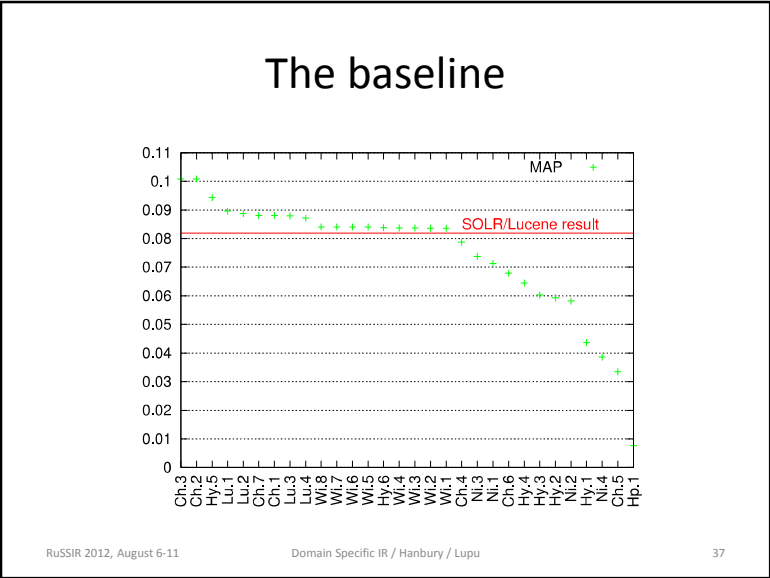
## Multilinguality

- Use the multilingual corpus to learn dictionaries
  - EN-JP [Nanba:2011]
  - “patentese” – EN [Nanba:2009]
    - Word processor = document processing device, document information processing device, document editing system, document writing support system
    - TV Camera = photographic device, image shooting apparatus, image pickup apparatus
  - In both cases, using hypernym-hyponym patterns in text

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

36



### Metadata

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization International Bureau

(43) International Publication Date 27 May 2004 (27.05.2004)

(10) International Publication Number WO 2004/043551 A1

<wo-patent-document id="example01" file="043551.xml"  
country="WO" doc-number="043551" kind="A1" date-published="20040527"  
dtd-version="v1.3 2005-01-01" lang="en">

<bibliographic-data id="bibl" country="WO" lang="en">

<publication-reference>  
<document-id>  
<country>WO</country>  
<doc-number>043551</doc-number>  
<kind>A1</kind>  
<date>20040527</date>  
</document-id>  
</publication-reference>

RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu 39

- ### Summary
- One can do a very decent job with a modern IR engine
  - Improvements come from
    - Splitting the query
    - Multi-word terms (sometimes)
  - Text analysis appears to be most useful in providing assistance to the user – through information extraction – rather than as an automated search process.
- RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu 38

- ### Kind codes
- Each office has its own kind codes
- | EPO   | USPTO  |
|---|--|
| • A1 APPLICATION PUBLISHED WITH SEARCH REPORT       | • A PATENT [FROM BEGIN UNTIL END 2000] or PATENT ISSUED AFTER 1ST PUB. WITHIN THE TVPP         |
| • A2 APPLICATION PUBLISHED WITHOUT SEARCH REPORT    | • A1 FIRST PUBLISHED PATENT APPLICATION [FROM 2001 ONWARDS]                                    |
| • A3 SEARCH REPORT                                  | • A2 REPUBLISHED PATENT APPLICATION [FROM 2001 ONWARDS]  |
| • A4 SUPPLEMENTARY SEARCH REPORT                    | • A9 CORRECTED PATENT APPLICATION [FROM 2001 ONWARDS]  |
| • A8 MODIFIED FIRST PAGE                            | • B1 REEXAM. CERTIF., N-ND REEXAM. or GRANTED PATENT AS FIRST PUBLICATION [FROM 2001 ONWARDS]  |
| • A9 MODIFIED COMPLETE SPECIFICATION                | • B2 REEXAM. CERTIF., N-ND REEXAM. or GRANTED PATENT AS SECOND PUBLICATION [FROM 2001 ONWARDS] |
| • B1 PATENT SPECIFICATION ( <i>granted patent</i> ) | • B3 REEXAM. CERTIF., N-ND REEXAM.   |
| • B2 NEW PATENT SPECIFICATION                       | • B8 CORRECTED FRONT PAGE GRANTED PATENT [FROM 2001 ONWARDS]                                   |
| • B3 AFTER LIMITATION PROCEDURE                     | • B9 CORRECTED COMPLETE GRANTED PATENT [FROM 2001 ONWARDS]                                     |
| • B8 MODIFIED FIRST PAGE GRANTED PATENT             |  |
| • B9 CORRECTED COMPLETE GRANTED PATENT              |  |
|   | • ...  |
- RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu 40

<classification-ipc id="ipc7">  
 <edition>7</edition>  
 <main-classification>A63B  
57/00</main-classification>  
</classification-ipc>  
  
<application-reference appl-type="PCT">  
 <document-id>  
 <country>GB</country>  
 <doc-number>004926</doc-number>  
 <date>20031113</date>  
 </document-id>  
</application-reference>  
  
<language-of-filing>en</language-of-filing>  
<language-of-publication>  
en  
</language-of-publication>  
</priority-claims>

(51) International Patent Classification?: A63B 57/00  
(21) International Application Number: PCT/GB2003/004926  
(22) International Filing Date: 13 November 2003 (13.11.2003)  
(25) Filing Language: English  
(26) Publication Language: English  
(30) Priority Data: 0226470.3 13 November 2002 (13.11.2002) GB  
  
<priority-claims>  
 <priority-claim>  
 <country>GB</country>  
 <doc-number>0226470.3</doc-number>  
 <date>20021113</date>  
 </priority-claim>  
</priority-claims>

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

41

Classification schemes

Office	Classification system
USPTO	*United States Patent Classification (USPC)
WIPO	International Patent Classification (IPC)
EPO	*European Classification (ECLA) – based on IPC, Indexing Codes (ICO)
JPO	File Index (FI) – based on IPC, Indexing Codes (F-terms)
KPO	IPC
SIPO	IPC

\* The USPTO and EPO will adopt, as of 2013, the Cooperative Patent Classification (CPC), which is based on ECLA/IPC

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

43

Patent classifications

- Patents are classified by the patent offices into large hierarchical classification schemes based on their area of technology
- Major benefits:
  - Access to concepts rather than words
  - Language independence
- Most classification is done manually by patent offices, although use of automated systems is increasing
- Classification schemes are regularly revised

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

42

IPC

• Sections:

Section	Description
A	Human necessities
B	Performing operations; Transporting
C	Chemistry; Metallurgy
D	Textiles
E	Fixed constructions
F	Mechanical engineering; Lighting, Heating, Weapons, Blasting
G	Physics
H	Electricity

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

44

NOT FOR REDISTRIBUTION

11

### IPC

- Example hierarchy:

Level	Number of categories	Example symbol	Example title
Section	8	G	Physics
Class	129	G04	Horology
Sub-class	631	G04D	Apparatus or tools specifically designed for making or maintaining clocks or watches
Main group	7392	G04D 3/00	Watchmakers' or watch-repairers' machines or tools for working materials
Sub-group	62493	G04D 3/04	Devices for placing bearing jewels, bearing sleeves, or the like in position

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

45

### Automated patent classification

- Has uses in patent offices for:
  - Pre-classification
  - Interactive classification
  - Re-classification
  - Promising application: classification of non-patent documents
- Common classification algorithms usually used: SVM, k-nearest neighbour, ...
- Recent classification tasks in the CLEF-IP and NTCIR Evaluation campaigns

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

47

### Characteristics of classification schemes

- Large imbalance in the distribution of documents in categories
- Most patents are assigned to multiple categories – a multi-classification task
- The codes are assigned at two levels of importance – primary categories and secondary categories

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

46

### Back to Meta-data

(72) Inventors; and  
(75) Inventors/Applicants (for US only): **THIRKETTLE, John, Scott** [GB/GB]; The Red House, 18 Station Road, Long Marston, Hertfordshire HP23 4QS (GB). **EMMERSON, Geoffrey** [GB/GB]; World Golf Systems Ltd, Axis 4 Rhodes Way, Watford, Herts, WD24 4YW (GB).

```
<parties>  
<applicants>  
  <applicant sequence="1" designation="all-except-us" app-type="applicant">  
    <addressbook>  
      <orgname>WORLD GOLF SYSTEMS LTD (GB)</orgname>  
      <address>  
        <street>Axis 4 Rhodes Way</street>  
        <city>Watford</city>  
        <county>Herts</county>  
        <postcode>WD24 4YW</postcode>  
        <country>GB</country>  
      </address>  
    </addressbook>  
  </applicant>  
</parties>
```

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

48

(72) Inventors; and  
(75) Inventors/Applicants (for US only): **THIRKETTLE, John, Scott** [GB/GB]; The Red House, 18 Station Road, Long Marston, Hertfordshire HP23 4QS (GB). **EMMERSON, Geoffrey** [GB/GB]; World Golf Systems Ltd, Axis 4 Rhodes Way, Watford, Herts, WD24 4YW (GB).

<parties>  
 <applicant sequence="2" designation="us-only" app-type="applicant-inventor">  
 <addressbook>  
 <last-name>THIRKETTLE</last-name>  
 <first-name>John</first-name>  
 <address>Somewhere over the rainbow</address>  
 </addressbook>  
 </applicant>  
 <applicant sequence="3" designation="us-only" app-type="applicant-inventor">  
 <addressbook>  
 <last-name>EMMERSON</last-name>  
 <first-name>Geoffrey</first-name>  
 <address>34 Ralph Waldo Pond</address>  
 </addressbook>  
 </applicant>  
</applicants>

RuSSIR 2012, August 6-11      Domain Specific IR / Hanbury / Lupu      49

INTERNATIONAL SEARCH REPORT

International Application No.  
PCT/us 03/04926

A. CLASSIFICATION OF SUBJECT MATTER  
IPC 7 A63B57/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED  
Minimum documentation searched (classification system followed by classification symbols)  
IPC 7 A63B

<search-report-data id="srep" lang="en" srep-type="isr" srep-office="EP">  
 <srep-for-pub>  
 <classification-ipc>  
 <edition>7</edition>  
 <main-classification>A63B 57/00</main-classification>  
 </classification-ipc>  
 <srep-fields-searched>  
 <minimum-documentation>  
 <classification-ipc>  
 <edition>7</edition>  
 <main-classification>A63B</main-classification>  
 </classification-ipc>  
 </minimum-documentation>  
 <database-searched>  
 <text>EPO internal, PAJ</text>  
 </database-searched>  
 </srep-fields-searched>

Fields searched  
(used)

RuSSIR 2012, August 6-11      Domain Specific IR / Hanbury / Lupu      51

(74) Agents: **POWELL, Stephen, David** et al.; Williams Powell, Morley House, 26-30 Holborn Viaduct, London EC1A 2BP (GB).

<agents>  
 <agent sequence="1" rep-type="agent">  
 <addressbook>  
 <last-name>POWELL</last-name>  
 <first-name>Stephen</first-name>  
 <middle-name>David</middle-name>  
 <suffix>et al</suffix>  
 <orgname>Williams Powell</orgname>  
 <address>  
 <building>Morley House</building>  
 <street>35 Kings Row</street>  
 </address>  
 </addressbook>  
 </agent>  
</agents>  
</parties>

RuSSIR 2012, August 6-11      Domain Specific IR / Hanbury / Lupu      50

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GB 2 364 924 A (HILLAN GRAHAM CARLYLE) 13 February 2002 (2002-02-13) page 5, line 22 -page 6, line 13; figures 1-4 abstract	1-11
X	US 5 248 144 A (ULLERICH SCOTT R) 28 September 1993 (1993-09-28) column 3, line 14 - line 68; figures 1-5	1-11

<srep-citations>  
 <citation>  
 <patcit dnum="GB2364924" id="sr-pcit0001" num="0001">  
 <document-id>  
 <country>GB</country>  
 <doc-number>2364924</doc-number>  
 <kind>A</kind>  
 <name>HILLAN GRAHAM CARLYLE</name>  
 <date>20020213</date>  
 </document-id>  
 <rel-passage>  
 <passage>page 5, line 22 - page 6, line 13; figures 1-4</passage>  
 <passage>abstract</passage>  
 </rel-passage>  
 <category>X</category>  
 <rel-claims>1-11</rel-claims>  
 </patcit>  
 </citation>  
 <citation>  
 <patcit dnum="US5248144" id="sr-pcit0002" num="0002">  
 <document-id>  
 <country>US</country>  
 <doc-number>5248144</doc-number>

RuSSIR 2012, August 6-11      Domain Specific IR / Hanbury / Lupu      52

### Pagerank (?)

RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu 53

### Classification

- Using classifications in ranking
- Classification was created to facilitate search
  - Manually
- How about automatically?

[Gobeil:2010]

Discarded field	MAP
Baseline	0.097
Title	0.096
Abstract	0.091
Claims	0.052
IPC 4-digits codes	0.0791
IPC complete codes	0.0842

RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu

Metric	Baseline	Using Classification Hierarchy
Num Patents Returned	93676	93676
Num Rel Patents Ret	1118	1121
MAP	0.0485	0.0626*
Recall@100	0.1888	0.2585*
nDCG	0.2245	0.2844*

[Harris:2011]

RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu 55

### Name disambiguation

- Or Synonym detection

IMPERIAL CHEMICAL INDUSTRIES PLC> IMPERIAL CHEMICAL INDUSTRIES PLC>ICI LTD 10039107  
FBC LIMITED> FBC LIMITED>FISONS LTD10177257  
ASSOCIATED ENGINEERING ITALY S.p.A.> ASSOCIATED ENGINEERING ITALY S.P.A.>ASS ENG ITALIA 10226032  
>BCIRA BRITISH CAST IRON RES ASS>BCIRA 10498172  
>NOVO NORDISK A/S NOVO INDUSTRI A/S>NOVO INDUSTRI AS 10498253  
>BICC Public Limited Company BRITISH INSULATED CALLENDERS>BICC PUBLIC LIMITED COMPANY 10498399  
DAVY MCKEE (OIL & CHEMICALS)LIMITED>DAVY MCKEE OIL & CHEM 10498706  
>BP Chemicals Limited BP CHEM INT LTD>BP CHEMICALS LIMITED 10502442  
>ENICHEM ELASTOMERS LIMITED >THE INTERNATIONAL SYNTHETIC RUBBER COMPANY LIMITED>ENICHEM ELASTOMERS  
>BRITISH TELECOMMUNICATIONS public limited company THE POST OFFICE>POST OFFICE 10504886  
S.A. SANOFI - LABAZ N.V.> S.A. LABAZ N.V.>LABAZ NV 10506339  
FORD-WERKE AKTIENGESSELLSCHAFT>FORD MOTOR COMPANY LIMITED>FORD MOTOR CO. 10507419  
>BASF Aktiengesellschaft NORSK HYDRO AS>NORSK HYDRO A.S.>NORSK HYDRO A/S 10507592  
International Business Machines Corporation> INTERNATIONAL BUSINESS MACHINES CORPORATION>IBM 10511969  
BAJ Limited> BAJ VICKERS LIMITED>BAJ VICKERS LTD 10514464  
>AstraZeneca AB>ZENECA LIMITED ICI PLC>ASTRAZENECA AB>IMPERIAL CHEMICAL INDUSTRIES PLC 10519727  
SCM CHEMICALS LIMITED> LAPORTE INDUSTRIES LIMITED>LAPORTE INDUSTRIES LTD 10521070  
Philips Electronics N.V.> N.V. PHILIPS' GLOEILAMPENFABRIEKEN>PHILIPS NV 10521825  
Procter & Gamble Limited> THE PROCTER & GAMBLE COMPANY>PROCTER & GAMBLE 10525897  
THE PROCTER & GAMBLE COMPANY> PROCTER & GAMBLE>P & G SPA 11411482  
>AVIO S.p.A>ELASIS SISTER CIERCA FIAT NEL M>AVIO S.P.A 8243658  
AVIO S.p.A>AVIO S P A>FIATAVIO SPA 11415073

RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu 54

### Citation analysis

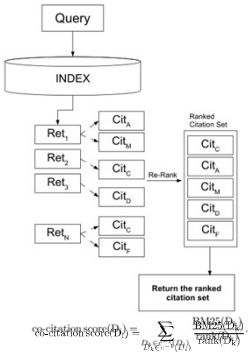
- Citations are used for
  - Evaluation
  - Boosting ranks
- First, a word of caution
  - In 1996, from all patents applied for at USPTO and EPO: 25% were granted only by the USPTO and 10% only by EPO [Michel:2001]

RuSSIR 2012, August 6-11 Domain Specific IR / Hanbury / Lupu 56

Citation analysis

[Gobeil:2009],[Gurulingappa:2010]

- Rerank the citations based on
  - Ranks of the documents citing them
  - Scores of the documents citing them



Data Sources

- Patent data
  - Patent offices
    - Rarely online, even more rarely bulk download
  - USPTO (via Google)
    - <http://www.google.com/googlebooks/uspto-patents.html>
  - Evaluation campaigns
    - Multi-office subsets

Citation analysis

- Promote patents that are cited by the retrieved patents [Gobeil:2010]
- Results improve drastically

	average length	average citations	IR (MAP)	IR + CitFB (MAP)
part 1	3860	32.5	0.045	0.208
part 2	6740	36.2	0.048	0.238
part 3	11800	47	0.048	0.282
part 4	33000	53.7	0.032	0.335
all the query set	13800	42.3	0.043	0.261

Figure 2. Comparison of the performances before (IR column) and after (IR + CitFB column) the citations feedback, regarding patents lengths.

- But not always:
  - same experiment in CLEF-IP showed much less improvement

The screenshot shows the Google USPTO Bulk Downloads: Patents page. It lists various patent data sources available for free download, categorized into Patent Grants, Patent Application Publications, and Additional Patent Data. The list includes items like Patent Grant Multi-Page Images, Patent Grant Full Text with Embedded Images, Patent Grant Bibliographic Data, Patent Grant OCR Text, Patent Grant Single-Page Images, PAIR (Patent Application Information Retrieval) Data, Patent Application Publication Multi-Page Images, Patent Application Publication Full Text with Embedded Images, Patent Application Publication Full Text, Patent Application Publication Bibliographic Data, Patent Application Single-Page Images, Patent Assignment Text, Patent Maintenance Fee Events, Patent Classification Information, and Patent IFW Petition Decisions.



## Data Sources

- Evaluation campaigns

NT CIR	Description	Approx. size
3	Japanese Patent Application fulltext 1998-1999 JAPIO Japanese abstracts (1995-1999) and PAJ English Abstract (1995-1999)	22GB
4	Japanese Patent Full-text 1993-1997, JPO English abstracts (1993-1997)	100GB
5	Japanese Patent Applications Full-text 1993-2002, JPO English abstracts (1993-2002)	100GB
6	NTCIR-5 + USPTO Patent grant data 1993-2002	152GB
7	NTCIR6 + scientific abstracts (EN and JP)	156GB
8	NTCIR7 + unexamined JP patent applications 1993-2007, patent grant data from USPTO 1993-2007	300GB
9	JP-EN and ZH-EN MT training data	10GB

RUSSIA 2012: August 6-11

Domain Specific IR / Hapbury / Lupu

61

## Data Sources

- EPO – Worldwide database
  - <https://data.epo.org/publication-server/>
  - DOCDB – master documentation database, with world-wide coverage

RuSSIR 2012, August 6-11

Domain Specific IR / Hanbury / Lupu

63

## Data Sources

- Evaluation campaigns

CLEF-IP	Description	Approx. size
2009	EP patent applications & grants 1985-2000	18GB
2010	EP patent applications & grants 1985-2001	19GB
2011	EP patent applications & grants 1985-2002 + WO documents referenced by the above EPO documents	15GB
TREC-CHEM	Description	Approx. size
2009	All USPTO, EPO, PAJ, WO publications until 2002, classified in IPC class C or A61K; Scientific Articles from the Royal Society of Chemistry	20GB
2010	TREC-CHEM 2009 + corresponding images, as well as scientific articles from Open Access Journals	420GB

RUSSIR 2012 August 6-11

Domain Specific IR / Hanbury / Lupu

62

## Data Sources

- EPO – Worldwide database
  - Open Patent Services (OPS)
  - Free resource of patent data, using a web-service interface
  - Fair use policy

```
protocol/authority/prefix/service/reference-type/  
input-format/input/[ endpoint] [ constituent(s) ] /  
output-format
```

The diagram shows the URL `http://ops.epo.org/2.6.2/rest-services/number-service/application/original/US.11380365.A1.20070515/docdb` with arrows pointing to specific parts and their semantic labels below:

- `http`: protocol
- `ops.epo.org`: authority
- `2.6.2`: version
- `rest-services`: prefix
- `number-service`: service
- `application`: reference type
- `original`: input format
- `US.11380365.A1`: CC
- `20070515`: number
- `docdb`: KC
- `64`: output format

Additional context at the bottom includes: "RuSSIR 2012, August 6-11" and "Domain Specific IR / Hanbury / Lupu".

RUSSIR 2012 August 6-11

Domain Specific IR / Hanbury / Lupu

64

## Example

- Fetch a full PDF

```
FullTextPDFClient ftpc = new FullTextPDFClient("EP", "0123456", "A2");
String filename = ftpc.getPdf();

public FullTextPDFClient(String country, String number, String kind) {
    this.country = country;
    this.number = number;
    this.kind = kind;
    String server = "http://ops.epo.org/2.6.2/rest-services/published-data/"
    BASE_URI = server + "publication/epodoc/" + country + number + "." + kind;
    com.sun.jersey.api.client.config.ClientConfig config = new com.sun.jersey.api.client.config.DefaultClientConfig();
    client = Client.create(config);
    imageInfo = client.resource(BASE_URI).path("images");
    BASE_URI = server + "images";
    pdfResource = client.resource(BASE_URI).path(country + "/" + number + "/" + kind + "/fullimage");
}
```



HANDBOOK ON INDUSTRIAL PROPERTY INFORMATION AND DOCUMENTATION

Ref.: Standards – ST.36      page: 3.36.2

STANDARD ST.36

Version 1.2

RECOMMENDATION FOR THE PROCESSING OF PATENT INFORMATION USING XML  
(EXTENSIBLE MARKUP LANGUAGE)

*Revision adopted by ST.36 Task Force of the Standards and Documentation  
Working Group (SDWG) on November 23, 2007*

## Example

```
public String getPdf() throws IOException, FileNotFoundException, ParserConfigurationException, SAXException {
    String ucid = country + "." + number + "." + kind;
    // get the information about this particular UCID.
    String opsData = imageInfo.accept("application/ops+xml").get(String.class);
    //process the info to find the number of pages
    int numberOfPages = getPathAndNumberOfPages(opsData);

    if (numberOfPages == 0) { return null; }

    //for each page, send a request to get it and save it in the temp folder
    for (int i = 1; i <= numberOfPages; i++) {
        BASE_URI = server + "images";
        if (path.contains("published-data")){
            path=path.replace("published-data/", "");
        }
        if (path.contains("images")){
            path=path.replace("images/", "");
        }
        pdfResource = client.resource(BASE_URI).path(path).queryParam("range", "" + i);
        ClientResponse cr = pdfResource.accept("application/pdf").get(ClientResponse.class);
        writePdfFile(cr, ucid + "-part" + i + ".pdf");
        System.out.println("Got page no. " + i);
    }
}
```

## A bit of history

- IR academic interest in Patent IR (formally) start:
  - Workshop on Patent Retrieval, SIGIR 2000
    - N. Kando and M.-K. Leong
  - Already introduces the key issues
    - Cross-lingual
    - Vocabulary
    - Explicit semantics
    - Interaction and visualization
    - evaluation

