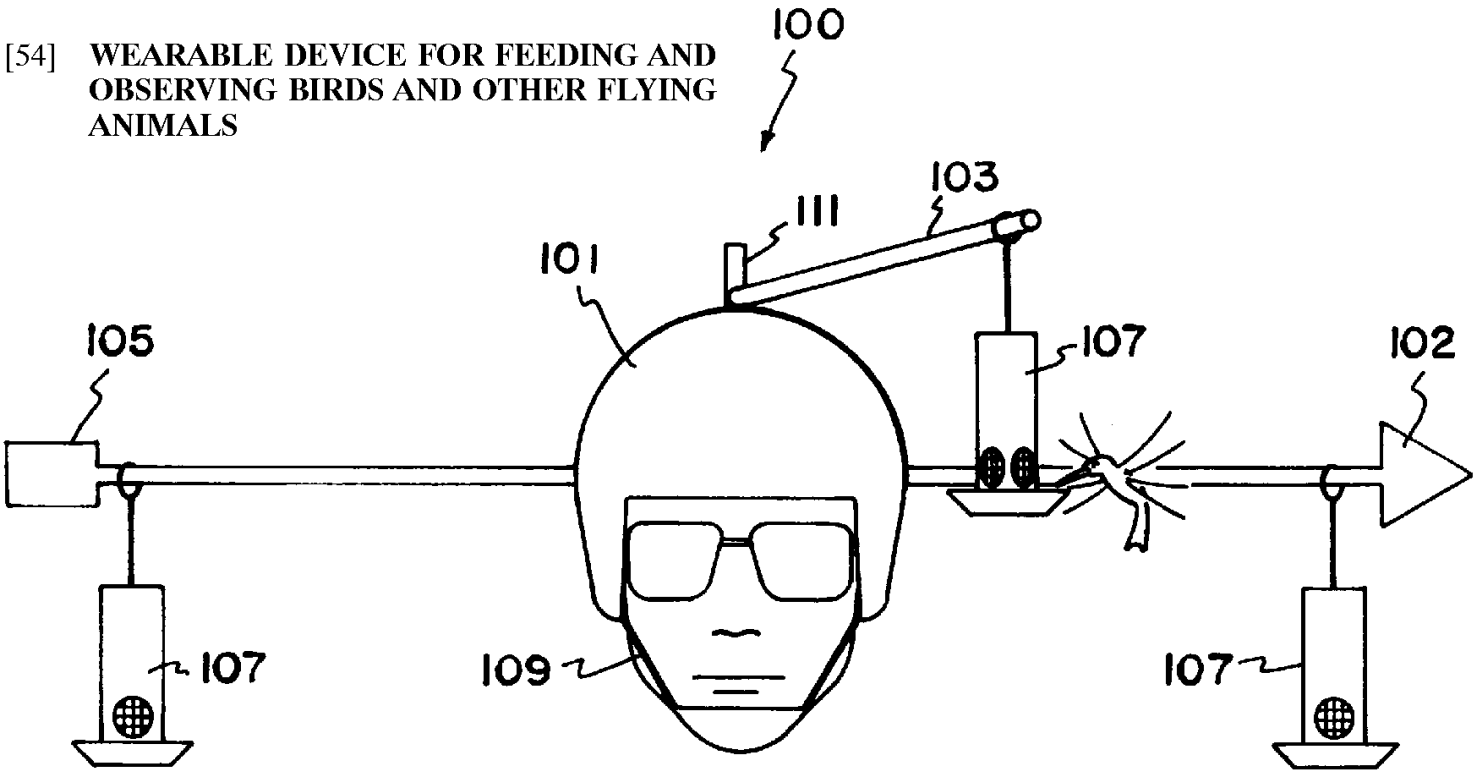
Challenges in Patent Retrieval and Mining

Dennis Hoppe

Bauhaus-Universität Weimar
dennis.hoppe@uni-weimar.de

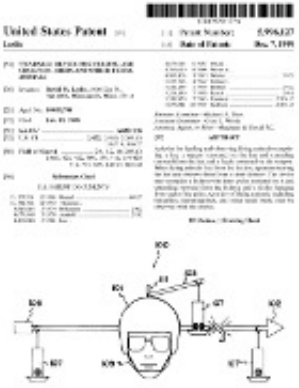
[54] **WEARABLE DEVICE FOR FEEDING AND OBSERVING BIRDS AND OTHER FLYING ANIMALS**

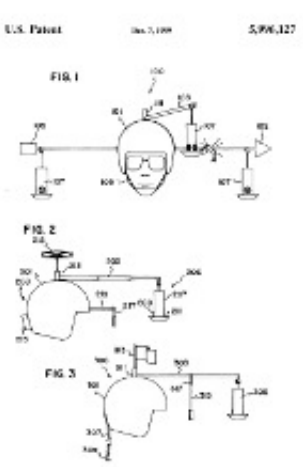


Challenges in Patent Retrieval and Mining

Patent Structure

- Multimedia data (drawings, mathematical formulas & chemical structures)
- Bibliometric data (classification codes, citations, . . .)
- Unstructured information (abstract, claims, description, . . .)





Challenges in Patent Retrieval and Mining

Research Areas

- Evaluation of patent retrieval
- Retrievability of patents
- Automated patent classification
- Image-based patent retrieval and classification
- Cross-language retrieval and classification
- Automated query generation
- Creation of technical trend-maps

Challenges in Patent Retrieval and Mining

Evaluation of Patent Retrieval

Motivation

- Precision-based evaluation measures do not meet challenges of recall-oriented domains like patent retrieval
- Evaluation still dominated by precision-based metrics, though

Challenges

- Only focussing on measuring the recall is not a solution
- Reward systems that return relevant patents earlier

PRES: A Score Metric for Evaluating Recall-Oriented IR Applications

[Magdy and Jones, 2010]

- Modification of the normalized recall measure [Rocchio, 1964]
- Takes the rank of relevant documents retrieved into account
- 0.87 correlation to recall, 0.66 correlation to precision
- Utilized at CLEF-IP 2010

	Ranks of rel. docs	Recall	F_1	F_4	PRES
System 1	{50, 51, 53, 54}	1	0.077	0.46	0.25
System 2	{1,2,3,4}	1	0.077	1	1
System 3	{1,98,99,100}	1	0.077	0.86	0.28

Challenges in Patent Retrieval and Mining

Retrievablity of Patents: Prior Art Search

Motivation

- Patent examiners need to determine the novelty of a new invention
- Search includes both patent and non-patent literature
- Even a single omitted patent can lead to significant financial losses

Challenges [Shinmori et al., 2003]

- Large vocabulary due to generic language
- Complex syntactic structure of patents
- Spelling errors, OCR errors
- Multilinguism

CLEF-IP: Prior Art Search task

- Find patents that are likely to outline prior art to a given patent application
- Multi-lingual data corpus (MAREC), over 2.6 million patents
- 67% English, 26% German, and 7% French patents

CLEF-IP: Prior Art Search task

- Find patents that are likely to outline prior art to a given patent application
- Multi-lingual data corpus (MAREC), over 2.6 million patents
- 67% English, 26% German, and 7% French patents

Literature Overview

- Retrieval Model Combination and Regression Models for Prior Art Search [Lopez and Romary, 2010]
- Improving Retrievability of Patents in Prior-Art Search [Bashir and Rauber, 2010]
- Preliminary Study into Query Translation for Patent Retrieval [Jochim et al., 2010]
- Patent Claim Decomposition for Improved Information Extraction [Parapatics and Dittenbach, PaIR 2009]
- Enhancing Patent Retrieval by Citation Analysis [Fuji, 2008]

Challenges in Patent Retrieval and Mining

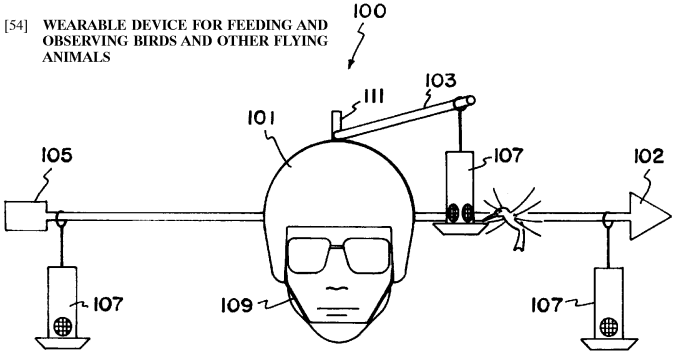
Automated Patent Classification

Motivation

- Organize the large amount of patent data
- Allow for a quick prior art search

International Patent Classification (IPC)

A	Section	Human necessities
42	Class	personal or domestic articles; headwear
B	Subclass	hats; head coverings
1/24	Group	hats [..] with means for attaching articles thereto



Automated Patent Classification

Challenges [Fall et al., 2003]

- Overlapping classes, e.g.,
  - C07 Organic chemistry
  - C08 Organic macromolecular compounds
- Unclear class boundaries, e.g.,
  - A42B Hats; head coverings
  - A41D Headbands; head-scarves
- Generic classes, e.g.,
  - G09 Educating; cryptography; display; advertising; seals
- Mapping of new technologies to the classification scheme is difficult

Automated Patent Classification

CLEF-IP: Classification task

- Classify a given patent document according to the IPC system, up to the subclass level

myClass: A Mature Tool for Patent Classification [Guyot et al., 2010]

- Single, multilingual classifier (Winnow algorithm)
- Features (unigrams, collocations) taken from a set of fields (inventor, applicant, title, abstract, claims, description)
- Language-specific stopwords removed
- Using collocations improves classification performance

Challenges in Patent Retrieval and Mining

Image-Based Patent Retrieval

Motivation

- Find patents relevant to a given image
- Improve the recall of a patent search

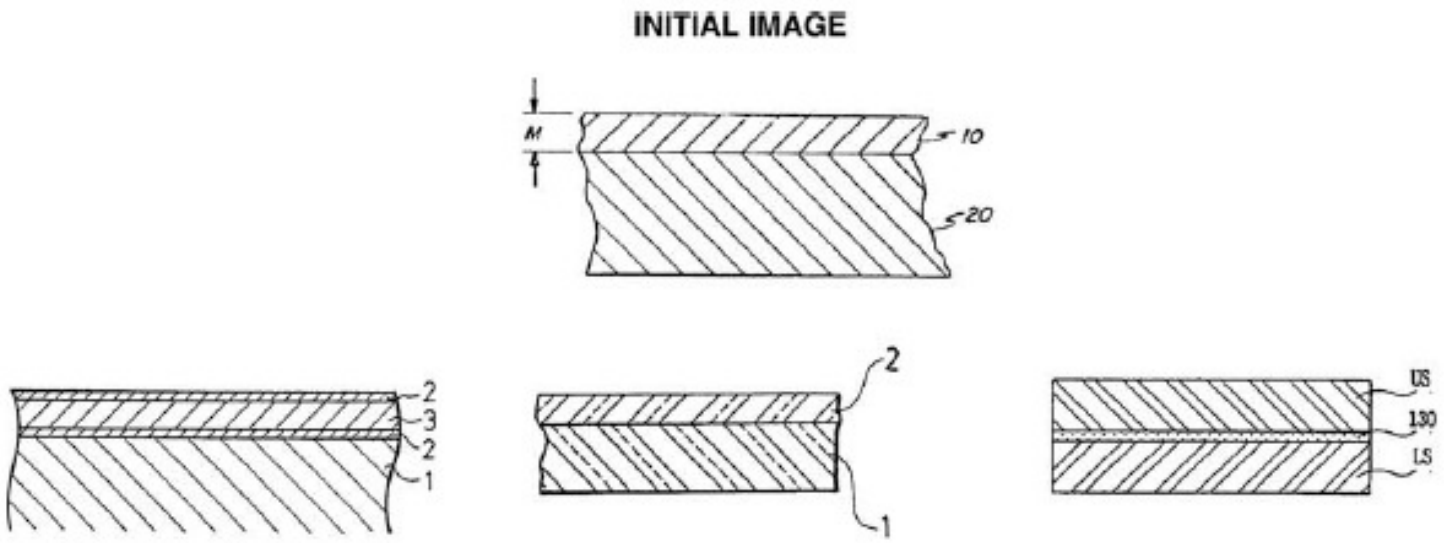
Challenges [Vrochidis et al., 2010]

- Patent drawings are black and white
- Hand-made sketches, rasterized images
- Scalability

Image-Based Patent Retrieval

Towards Content-Based Patent Image Retrieval [Vrochidis et al., 2010]

- Visual feature extraction based on Yang et al., 2006
- Exploit the description of figures



Challenges in Patent Retrieval and Mining

Towards an Ideal Patent Retrieval System

Challenges in Patent Retrieval and Mining

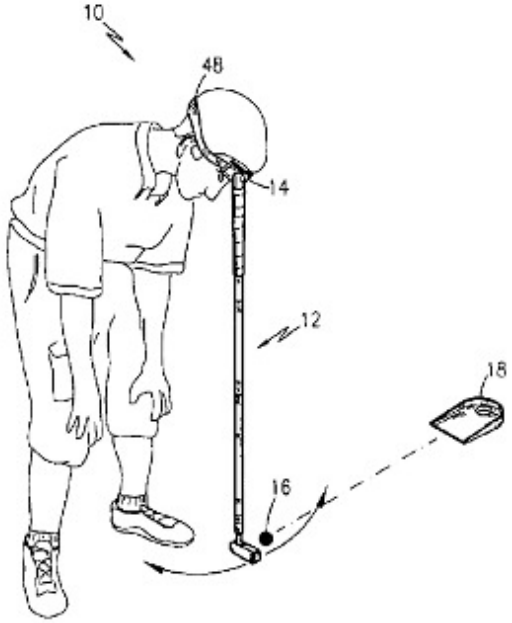
Towards an Ideal Patent Retrieval System

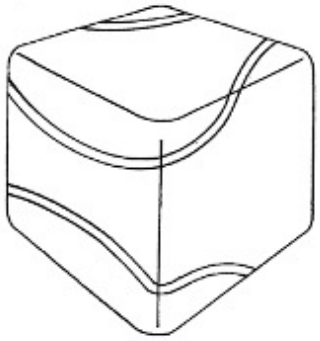
- Access to all patents worldwide
- Automatic translation of patents
- Support searching for similar drawings
- Citation-based traversal

[Simmons, E.: Patent Databases and Gresham's Law. WPI 2006]

Challenges in Patent Retrieval and Mining

Thank you





Challenges in Patent Retrieval and Mining

References

- Patent Claim Processing for Readability: Structure Analysis and Term Explanation.
  [Shinmori, A., Okumura, M., Marukawa, Y., and Iwayama M., ACL Workshop Proceedings, 2003]
- Multiple Retrieval Models and Regression Models for Prior Art Search.
  [Lopez, P. and Romary, L., CLEF-IP, 2010]
- Automated Categorization in the International Patent Classification.
  [Fall, C.J., Torcsvari, A., Benzineb, K., and Karetka, G., ACM SIGIR Forum, 2003]
- myClass: A Mature Tool for Patent Classification.
  [Guyot, J., Benzineb, K., and Falquet, G., CLEF-IP, 2010]
- Towards Content-based Patent Image Retrieval; A Framework Perspective.
  [Papadopoulos, S., Moumtzidou, A., Sidiropoulos, P., Pianta, E., Kompatsiaris, I., WPI Journal, 2010]
- PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications.
  [Magdy, W. and Jones, G. J.F., SIGIR, 2010]