Automatic classification of patents oriented to TRIZ: *a case study on large aperture optical elements*

Zhengyin Hu, Shu Fang, Wen Yi, Xian Zhang,Tian Liang

Chengdu Document and Information Center ,

Chinese Academy of Sciences

Leiden, The Netherlands
Sept .2, 2014

Contents

1. Introduction

Research & Applications of Intellectual Property in CAS

- **Information Portal:** IP database, IP analysis tools, IP training, IP assessment, IP trading & transforming, etc.
- **Intelligence products:** IP rights information journal, IP analysis reports, IP consulting reports, Patents Tech Mining, etc.
- **Services:** custom data, intelligence products, consulting, training services for researchers, IP managers, IP policymakers, etc.
- **Groups:** IP management department of CAS, *IP Services group of CAS*, IP assistants in research institutes of CAS.

Research & Applications of Intellectual Property in CAS

文献、网络资
源、政策、新

战略先导科技专项知识产权分析（系列）

重点领域专题知识产权情报研究（系列）

Classification Schema based on Patent Code

- **International Patent Classification (IPC):** 8 Sections, ~69,000 classes
- **US Patent Codes:** 3 Groups, 462 Categories, ~153,000 classes
- **EPO Cooperative Patent Classification:** extension of the IPC, adding 18,400 refined subclasses.

A	Human necessities
B	Performing operations, transporting
C	Chemistry, metallurgy
D	Textiles, paper
E	Fixed constructions
F	Mechanical engineering, lighting, heating, weapons, blasting
G	Physics
H	Electricity

Classifications Schema based on Contradictions & Principles

- **TRIZ:** Russian acronym for Inventive Problem Solving Theory
- **Contradictions (Problems):** basic and common problems in one area. 1201 standard engineering problems were summarized.
- **Principles (Solutions):** basic and common solutions used for these problems. 40 Inventive Principles were summarized.

Inventive Principles
1. Segmentation
2. Extraction, Separation, Removal, Segregation
3. Local Quality
4. Asymmetry
5. Combining, Integration, Merging
6. Universality, Multi-functionality
7. Nesting
8. Counterweight, Levitation
9. Preliminary anti-action, Prior counteraction
10. Prior action

Advantage & Disadvantage of two classification schemas

- **Schema based on Patent Code :**

  Advantage: *mature; focused on technology field*

  Disadvantage: *stable and kept invariant for a long time;*
  too general to represent specific tech

- **Schema based on Contradictions & Principles :**

  Advantage: *mature; focused on similar problems &solutions*

  Disadvantage: *stable and kept invariant for a long time;*
  focused on machinery patents

Personalized Classifications Schema oriented to TRIZ

- **Dynamic Schema:** *from specific patents set, more accurate with more details*

- **Oriented to Problems & Solutions (P&S):** *help find patents with similar problems or solutions*

- **Rich Semantic Knowledge Representation (SKR):** *support deep tech mining on patents*

2. Methodology

Construct Classification Schema

- Micro-Level(SAO)
- Meso-Level(P&S)
- Macro-Level(Tech)

Preliminarily Classify Patents

- Features Selection
- Algorithms Selection
- Compare Classifiers

Optimize Classifier

- Smooth Imbalanced Data
- Reduce Dimension of SAO

Construct Classification Schema

- **Micro-Level(SAO Semantic Units):** extract Subject-Action-Object(SAO) triples from fields, such as Title, Abstract and clean SAO using *Term Clumping*.

  Results: patents are represented as bag-of-SAO.

  Tools: Relationship Extract Tool: Reverb,

  Text Analysis Software: Thomson Data Analyzer(VantagePoint)

- **Meso-Level(P&S Topics):** generate P&S topics based on bag-of-SAO of patents using LDA topic model.

  Results: patent-P&S matrix, P&S-SAO matrix.

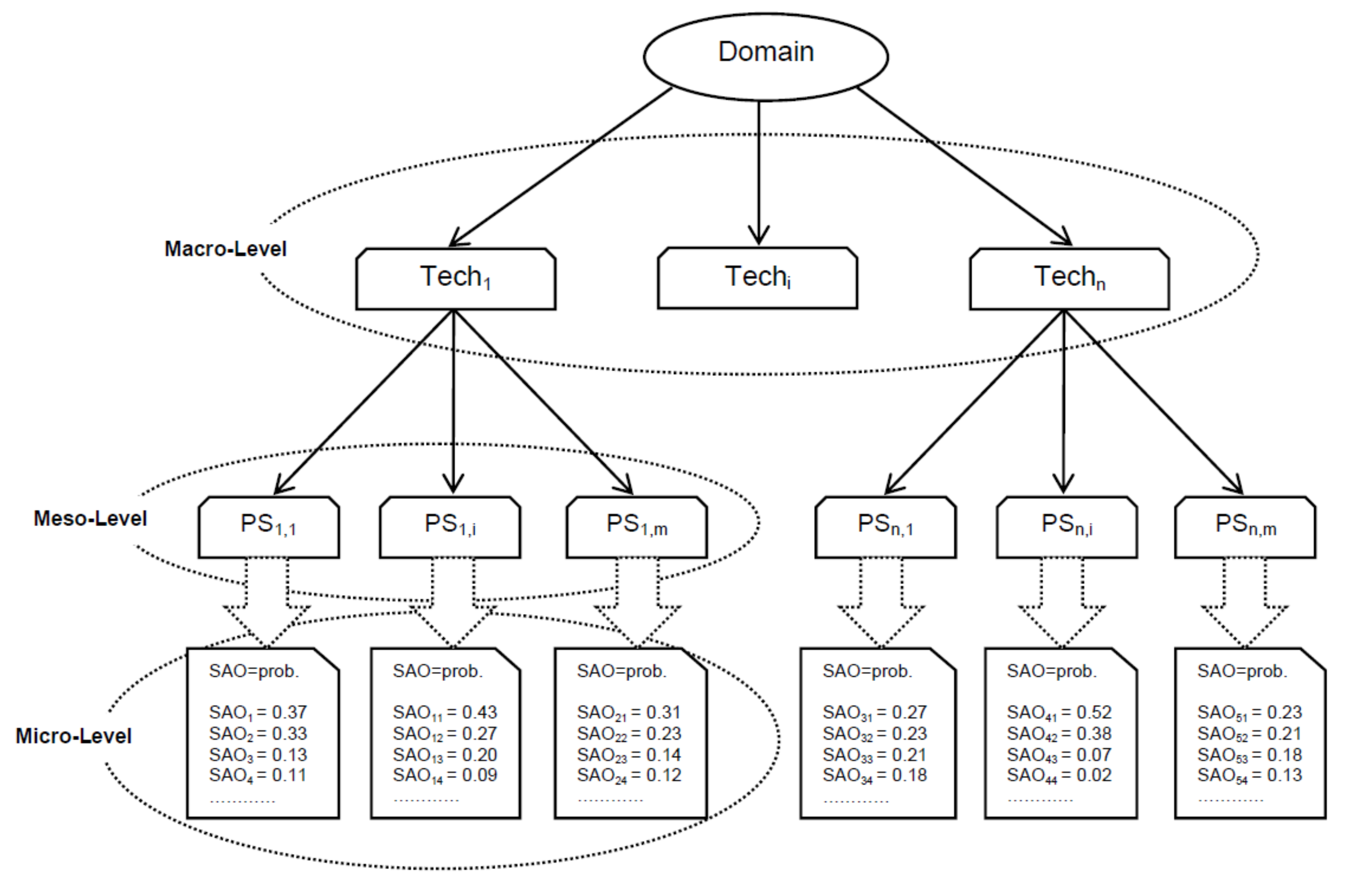
  Tools: Machine Learning Toolkit : MALLET

- **Macro-Level(Tech Topics):** generate Tech topics based on patents-P&S matrix using LDA topic model.

  Results: patent-Tech matrix, Tech-P&S matrix.

Construct Classification Schema

- **Experts:** prune meaningless topics, summarize similar topics and attach labels to topics.



Preliminarily Classify Patents

- **Feature Selection:** Information Gain (IG),Document Frequency (DF)

- **Classification Algorithms :** Maximum Entropy Classifier (MaxEnt), C4.5 Decision Tree Classifier (DT), Naïve Bayes(NB)

- **Compare Classifiers:** compare the accuracy based on the different combinations of features and algorithms and choose ***best combination*** to preliminarily classify patents on Test Sets.

Optimize Classifier

- **Smooth Imbalanced Data :** optimize the training set by over-sampling.
- **Reduce Dimension of SAO:** merge SAO by pattern rules to reduce dimensions of SAO features.
- **Build a new classifier:** apply the chosen combination of feature and algorithm on a new training set and SAO feature set.

3. Case Study

Construct Classification Schema

- **Data set:** choose Large Aperture Optical Elements (LAOE) patents as case study and get 1364 patents from Derwent Innovations Index(DII).

- **Micro-Level(SAO):** *2372* SAO were collected as the micro-level of the schema and patents were represented as bag-of-SAO.

- **Meso-Level(P&S Topics):** *200* P&S topics based on patents-SAO matrix were generated and experts chose *124* meaningful P&S topics as the meso-level of the schema.

- **Macro-Level(Tech Topics):** *20* Tech topics based on patents-P&S matrix were generated and experts summarized 4 Tech domain topics as the macro-level of the schema.

Construct Classification Schema

- Part of the personalized LAOE patent classification schema

Class No.	Tech Domains	P&S Topics	SAO semantic units
C1	Measuring surface shape	$P\&S_{10}(p=0.557)$	**check** large lens convex surface;
		$P\&S_{11}(p=0.213)$	**measure** surface roughness;
		$P\&S_{14}(p=0.117)$	**analyze** object surface profile;
		......	....
C2	Surface measuring method	$P\&S_{1}(p=0.628)$	method **measure** diffraction;
		$P\&S_{39}(p=0.124)$	method **measure** optical curvature;
		$P\&S_{114}(p=0.017)$	method **analyze** interference-fringe;
		......	....
C3	Surface measuring device	$P\&S_{15}(p=0.415)$	device **measure** wave aberration;
		$P\&S_{79}(p=0.354)$	**shear** interferometer for flatness;
		$P\&S_{102}(p=0.203)$	device **measure** lens deflection;
		......	....
C4	Online monitoring	$P\&S_{27}(p=0.813)$	**monitor** surface quality;
		$P\&S_{42}(p=0.102)$	**control** optical surface quality;
		$P\&S_{78}(p=0.005)$	**inspect** surface shape;
		......	....

Preliminarily Classify Patents

- We choose ***100*** patents as training set. And experts manually classify these patents to {*C1, C2, C3, C4*} as the training set.

- **Feature Selection:** top 5,10 and 20 IG SAO;
  DF above the threshold 2,3 and 5 SAO;

- **Algorithms Selection:** Maximum Entropy Classifier (MaxEnt), C4.5 Decision Tree Classifier (DT), Naïve Bayes(NB) in Mallet

Accuracy of Classifiers on Training Set

Feature Selection	MaxEnt(%)	DT(%)	NB(%)
IG (top5)	67.6%	74.6%	72.6%
IG (top10)	73.5%	*82.8%*	80.5%
IG (top20)	71.3%	77.2%	79.1%
DF(threshold=2)	57.2%	65.2%	71.2%
DF(threshold=3)	52.6%	68.9%	69.3%
DF(threshold=5)	59.3%	72.7%	74.8%
DF(threshold=3)	52.6%	68.9%	69.3%

Average Classification Results on 3 Test Sets

Class No	Precision	Recall	*F*-measure
C1	0.764	0.72	0.741
C2	0.792	0.64	0.708
C3	0.832	0.78	0.805
C4	0.718	0.72	0.719
{C1,C2,C3,C4}	0.784	0.72	0.743

Average Classification Results on Test Sets after Optimization

- **Optimization Strategy:** over-sampling to build a new training set and merging SAO to build a new classifier

Class No	Precision	Recall	*F*-measure
C1	0.884	0.82	0.851
C2	0.926	0.88	0.902
C3	0.782	0.78	0.781
C4	0.726	0.86	0.787
{C1,C2,C3,C4}	0.830	0.84	0.830

Result

- The SAO triples are more suitable as basic semantic units than keywords in patent tech mining oriented to TRIZ .

- Topic model can help mine P&S topics from SAO triples and Tech domains from P&S topics.

- The personalized classification schemes oriented to TRIZ can help deep patent tech mining.

- The dimension reduction of SAO based on pattern rules is important to the results of classification.

Discussion

- It is a challenge to automatically distinguish the Problems or Solutions from the topics generated on SAO triples.

- Less SAO is good for better feature selection, but is not good for topic model. There are two different SAO clumping standards for topic model and feature selection .

- The personalized classification schemes can be used as semantic index. How to apply it for other applications?

Thank You !