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

Zhengyin Hu*, Shu Fang and Tian Liang

*huzy@clas.ac.cn
Chengdu Document and Information Center, Chinese Academy of Sciences (China)

## Introduction

Most of the current available patent classification systems are based on patent classification code such as IPC, which are oriented to technology fields and not suitable for TRIZ users who pay more attention to find analogous patents in other fields that have solved the similar technical problems by using the same solutions (namely, TRIZ principles) (He 2007). The TRIZ principles can be considered as a patent classification schema. However, they are abstract and mainly focus on mechanism patents, which is hard to be directly applied to other domains. It will be more helpful for TRIZ users to classify patents according to a more agile, detailed schema by combining text mining and TRIZ principles.

## Methodology

In this paper, we propose an approach to automatically classify patents oriented to TRIZ applications based on a personalized classification schema. Firstly, we construct a personalized classification schema in micro-meso-macro levels. Micro-level is composed of Subject-Action-Object (SAO) extracted from patent text, meso-level is composed of Problems solved and Solutions used (P&S) topics generated by Latent Dirichlet Allocation (LDA) topic model based on SAO (Mimno 2011) and macro-level is composed of technology domain generated by LDA based on P&S topics. By consulting with experts, the final personalized patent classification schema for TRIZ is constructed. Then, we choose an appropriate feature and classifier to preliminarily classify patents according to the personalized classification schema. Finally, the classifier is optimized by smoothing imbalanced data and reducing features dimensions of SAO.

## Results

Large Aperture Optical Elements (LAOE) patents were selected as a case study. We selected the database of DII as data source and obtained 1364 patent documents covering LAOE domains. A relationship extraction tool named Reverb was used to extract raw SAO from the "Title" and "Abstract" fields. A text mining tool, VantagePoint, and domain thesauri were used to carry out SAO clumping (Zhang et al.2011). Then 2372 SAO were collected as the micro-level of the classification schema and the reference corpus of the bag-of-SAO.

The LDA module in MALLET, a machine learning tools set, was applied to generate 200 raw P&S topics based on SAO. Experts chose 124 P&S topics which were manually attached labels as the meso-level of the classification schema. We applied LDA on the 200 raw P&S topics to generate 20 upper technology domains and summarized 4 domains as the macro-level of the classification schema. Then we constructed a personalized patent classification schema for TRIZ which include 4 technology domains, 124 P&S topics and 2372 SAO. Part of the schema is indicated in Table 1.

Next, by comparing the accuracy rates which were computed by different combinations of features and classifiers, we chose top 10 of SAO information gain as the classification features and C4.5 decision tree as the classifier to preliminarily classify patents.

Finally, we optimized the classifier. Considering the dataset is medium, over-sampling approach was chose to deal with the imbalance in the data set. By analyzing the results of LDA, we induced some pattern rules of SAO. Then we merged them by these rules to reduce dimensions of SAO features. After that, we classified patents based on a new training set and SAO feature set and got a better result indicated in Table 2.

**Discussion and Conclusions**

This approach implements semi-automatic construction of a personalized classification schema using topic model and automatic classification of patents oriented to TRIZ applications. In medium size data set, this approach can classify patents with high accuracy and speed. The personalized classification schemes can facilitate TRIZ users to deeply utilize patents, too.

However, there are some challenges. We use SAO to generate P&S topics, but sometimes the topics are difficult to be induced into valid or meaningful P&S. It is another challenge to automatically distinguish the problems or solutions topics. And this approach is not available in small size data set (several hundred) and not verified in big size data set (tens of thousands).

**Tables**

Table 1. 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; |
| | | ..... | ..... |

Table 2. the final classify quality of LAOE patents

| Class No. | Precision | Recall | F-measure(F1) |
|---|---|---|---|
| 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 |

**References**

He, C. (2007). Automatic patent classification according to the 40 TRIZ inventive principles. Ph.D. Dissertation. National University of Singapore, pp.106-111.

Mimno, D. (2011). Machine Learning with MALLET. Resource document. Information Extraction and Synthesis Laboratory, Department of CS UMass, Amherst. http://mallet.cs.umass.edu/mallet-tutorial.pdf. Accessed 20 April 2014.

Zhang, Y., Porter, A. L., Hu, Z. Y. , Guo, Y.& Newman, N.C.(2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells, *Technological Forecasting and Social Change*, available online 28 January 2014.