Empirical study of constructing knowledge organization system of patent documents using topic model

Zhengyin Hu (huzy@clas.ac.cn Chengdu Library of the Chinese Academy of Sciences), Shu Fang (fangsh@clas.ac.cn Chengdu Library, CAS), Xian Zhang (zhangx@clas.ac.cn Chengdu Library ,CAS), Tian Liang (liangt @clas.ac.cn Chengdu Library, CAS)

Yi Zhang (yi.zhang.bit@gmail.com School of Management and Economics, Beijing Institute of Technology)

Summary

We use three methods to construct knowledge organization system (KOS) of patent documents. One is KeyGraph, one is Latent Dirichlet Allocation (LDA) model, and another is Hierarchical Latent Dirichlet Allocation (hLDA) model.

KeyGraph is a novel and efficient method for topic detection. It applies graph analytical methods to discover topics and their features. LDA is a classic topic model method for automatically organizing, understanding, and summarizing text corpora based on hierarchical Bayesian model. Hierarchical Latent Dirichlet Allocation is an extension of LDA. It can not only extract autonomously the topics, but also compute the hierarchical structure of them. It arranges the topics into a tree, with the desideratum that more general topics should appear near the root and more specialized topics near the leaves.

We selected the field of graphene sensor for empirical analysis, and compute the topics distributions of technology terms.

Methodology

- Firstly, build a set of patent documents.
- Secondly, extract terms of technology and conduct term clumping for topic model.
- Thirdly, compute terms graphs and get topics of terms using KeyGraph algorithm (Sayyadi, Hurst & Maykov, 2011).
- Fourthly, use LDA model to get topics of terms (David M., 2012).
- Fifthly, use hLDA model to get hierarchical topics of terms (David M., 2012).
- Lastly, compare the results of the three methods.

Empirical Analysis

- Firstly, we selected the database of Derwent Innovations Index for analysis and invited experts to determine the patent retrieval strategy and got 251 patent documents.
- Secondly, we extracted the keywords from the patent text fields and conduct term clumping to clean the technology terms of graphene sensor using Thomson Data Analyser (TDA). And we got 6658 technology terms.
- Thirdly, we used KeyGraph algorithm for topic model and got 128 topics by giving appropriate values to the params of KeyGraph algorithm.
- Fourthly, we used LDA model to get 10 topics of terms and every topic include 10 terms.

Method	TOPIC No	Topic Terms
KeyGraph	1	memory,manganese,manufacturing method, magnetic detection device,lcd,motor vehicle,nitrogen,nickel,monolayer,magnetization
	2	magnesium, low density,magneto resistance sensor, lower surface
	3	oxidation,proteins,oligonucleotide,molecular sensor,nanopore,mixing,nucleotide
	4	pixels,perpendicular,operation,polyethylene terephthalate, probe molecule
	5	optical sensor,primary,pollutants,protecting human eye, optical limiting response
LDA	1	transponder antennas, second container 3, source electrode 131,back-gate substrate, semi-conductive, groove 11,interplanar, large area scale, ink-jet printing, ultraviolet
	2	Initiator, broad range, west directions, organic-material device, turn, high strength mechanical property, one form, translocation, thin film magnetic recording, Raman measurements
	3	molecular spacer, carbon source-containing gas 24,Several conductive nanoparticles, terminal extension, continuous process, normal atmospheric conditions, top basal plane, dispersion medium comprising sheets, inorganic oxide particles, radio frequency interference
	4	antennae assembly electrode comprising, centrifugation system, doped semiconductor materials, high velocity airflow, graphene nanopore sensor 10, graphene platelets, stabilized graphene suspension, fuel filter clamps, horse, packer element
	5	smart gels, layer opening 17, uniform neck width, Hybrid sensor array, nanogap arrays increases, passive thermocouple, pure acetone, high electric field 14, 1 method, core polymer

Fig 1. Part of results of KeyGraph and LDA

- Fifthly, we used hLDA model to get hierarchical topics of terms and the depth of tree is 3.

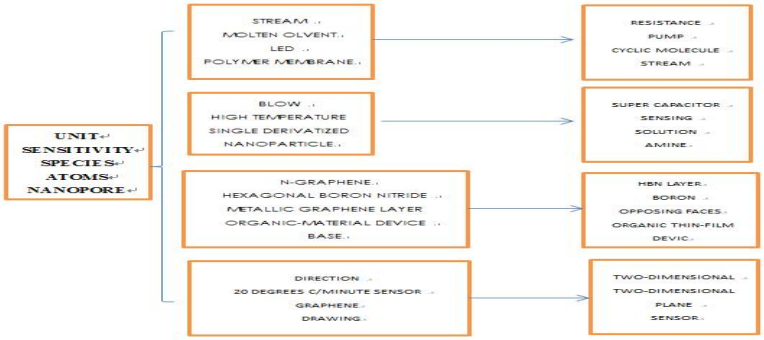

Fig 2. Part of result of hLDA

- Finally, we compared the results of the three methods. And the result shows that hLDA is a good tool for constructing KOS.

Further Work

- Firstly, we will try to get better result of topic model by optimizing the params setting of hLDA.
- Secondly, we will try to mapping the results of hLDA to the standard KOS, for example SKOS, etc.
- Finally, we will try to using RDF to describe the results of hLDA and storing them into knowledge base for further applications.

Part of References

- H. Sayyadi, M. Hurst, and A. Maykov. “KeyGraph: A Graph Analytical Approach For Fast Topic Detection”. http://keygraph.codeplex.com/. (2011-10-10).
- David M. Blei. “Latent Dirichlet allocation(LDA)”. http://www.cs.princeton.edu/~blei/lda-c/index.html. (2012-03-10).
- David M. Blei. “Hierarchical latent Dirichlet allocation(hLDA)”. http://www.cs.princeton.edu/~blei/downloads/hlda-c.tgz. (2012-03-20).