Patent Landscapes & New Technology Trends in IoT: Extracting and Visualizing Data Patterns

Nikolay Stulov
Department of Control and Applied Mathematics
Moscow Institute of Physics and technology
Moscow, Russia
nikolay.stulov@phystech.edu

***Abstract*—Extracting complex relations in unstructured data is a challenging and promising task in any field, especially fast-growing like Internet of Things (IoT). In this work we research different methods to extract and represent these relations. As a result, we present a set of text mining and patent mining tools and an approach to further building knowledge-based decision support system.**

***Keywords*—patent mining, text mining, ontology, decision support, patent landscape, information extraction, ontology driven information extraction, Internet of Things**

I. Introduction

A. Objective and relevance

In this work we research the mechanism of patent analysis and propose an intelligent system for this task. Formally, the problems are stated as follows:

- Extract and store knowledge of a specific field from a set of partially structured documents.
- Analyze and compare tools for visual representation of a subset of this knowledge from model.
- Propose a function to calculate probabilities of existence or appearance of model-unknown connections.

B. Existing Approaches

Currently a huge part of patent analysis in industry is done manually with the help of experts. Some approaches are known to automate this task using co-citation analysis [1], [2], [3]. Other approaches include ontology driven analysis as in [4], [5]. This work continues and improves discussed approach by taking implicit linguistic data from abstracts, claims and full texts into consideration.

Visualization of big bibliographic networks is usually done with the help of Visualization of Similarities (VOS) [6] algorithm, which is basically specifically weighted Multidimensional Scaling (MDS) [7]. This work proposes a number of different ways to set weights and compares them to VOS.

The rest of the paper is organized as follows. The next section describes steps of proposed approach in detail. Also, some examples and visualizations are given. After that there is evaluation section where analysis results are discussed. We conclude with further research review and comparison with above stated analogies.

II. Our approach detailed

A. Semantic-aware knowledge extraction

Since the main feature of our approach is implicit linguistic information retrieval, this step requires usage of semantic technologies. In this case we use Part of Speech (POS) tagging and semantic features of patent genre to make use of different entities that appear in the texts. Extraction was done as follows. Pre-structured data is extracted from XML as-is, but abstract and full text are treated individually. With POS tagging we filter out only noun groups (NG) of two words (bigrams). After that the sentences are encoded by the number of occurrences of each NG, a matrix representation X of text is built. Then we use outer product on X to build first order collocation matrix T, which is then scaled with Term Frequency — Inverse Document Frequency (TF-IDF) scale. Finally we run PageRank on T and choose $n = 100$ highest ranked noun groups as output (keywords).

B. Ontology

To represent knowledge we designed an ontology of our field, which includes entities "Patent", "Author", "Assignee", "Region", "Class" and "Keyword", trivial accessory relations, a citing relation and analogy relation between two patents. For implementation purposes graph database was used. The graph is very complicated due to the number of entities, as expected. For example, the citation network is depicted in Fig. 1.

Some interesting analysis can be applied at this point already. Sorting out authors who have more than one patent (active authors) leads to obvious clustering (see Fig. 2), which appears to be regional first (see Fig. 3).

Some predictions can also be made. As Fig. 3 shows, patents with unavailable regional information can be assigned to a certain region.

C. Building landscapes

Images above only depict subsets of ontology graph and posses no information about likeness between entities. To build landscapes means to project multidimensional data into two-dimensional. There are many methods to tackle this problem.

The MDS approach optimizes loss function:

$$\mathcal{L}(\mathbf{x}_1, ..., \mathbf{x}_n) = \frac{\sum\limits_{i<j} w_{ij} \left(f(p_{ij}) - ||\mathbf{x}_i - \mathbf{x}_j||_2 \right)}{\sum\limits_{i<j} w_{ij} f(p_{ij})^2} \qquad (1)$$

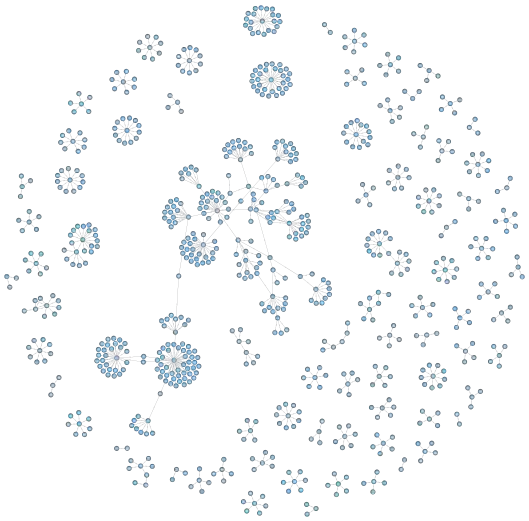

Figure 1. Patents (blue) co-citation network.

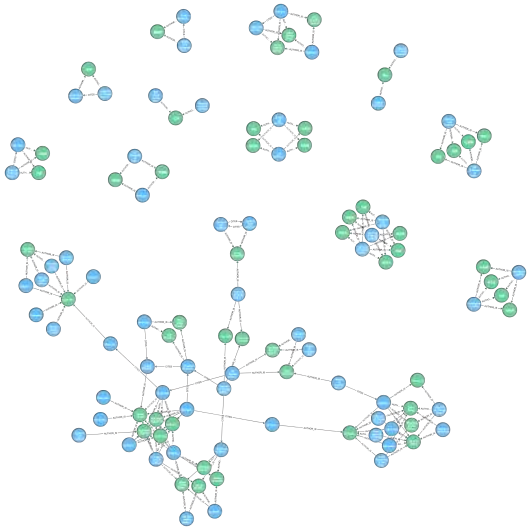

Figure 2. Patents (blue) and active authors (green) network.

where w_{ij} are weights, f denotes transformation function of proximity values p_{ij}. Usually the weights w_{ij} are set to 1.

In [7] they show that VOS solution is equivalent to MDS solution with $p_{ij} = \frac{1}{s_{ij}}$ and $w_{ij} = s_{ij}$, where

$$s_{ij} = \frac{2mc_{ij}}{c_i c_j} \tag{2}$$

where c_i denotes the total number of links of node i and m denotes the total number of links in the network.

This approach proved good at building bibliographic maps. But as soon as network contains information of different types from different sources, it becomes possible to use other metrics more efficiently. To do this we need to vectorize objects in some way. We propose using euclidean distance (ED), cosine similarity and traingle's area similarity - sector's area similarity (TS-SS) [8] on concatenation of attribute vectors.

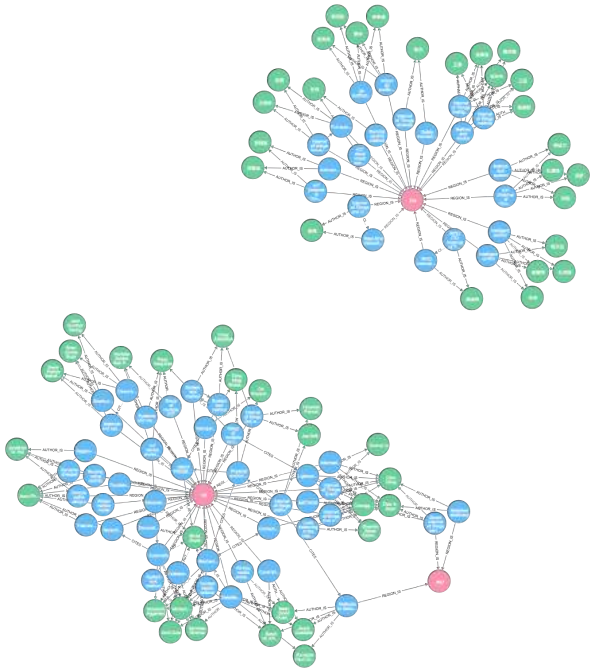

Figure 3. Regional subnetworks.

Cosine similarity is given by:

$$V = \text{cosine}(\mathbf{a}, \mathbf{b}) = \frac{(\mathbf{a} \cdot \mathbf{b})}{||\mathbf{a}||_2 \cdot ||\mathbf{b}||_2} \tag{3}$$

Euclidean distance (ED) is given by:

$$\text{ED}(\mathbf{a}, \mathbf{b}) = ||\mathbf{a} - \mathbf{b}||_2 = \sqrt{\sum_{i=0}^{n} (a_i - b_i)^2} \tag{4}$$

TS-SS is given by:

$$\text{TS-SS}(\mathbf{a}, \mathbf{b}) = \frac{\pi \sin(\theta')\theta'}{720} ||\mathbf{a}||_2 ||\mathbf{b}||_2 \cdot \cdot \left(\text{ED}(\mathbf{a}, \mathbf{b}) + \text{MD}(\mathbf{a}, \mathbf{b})\right)^2 \tag{5}$$

where Magnitude distance (MD) is:

$$\text{MD}(\mathbf{a}, \mathbf{b}) = \left| \sqrt{\sum_{i=0}^{n} a_i^2} - \sqrt{\sum_{i=0}^{n} b_i^2} \right| \tag{6}$$

In most trivial cases attribute vectors are one-hot encoded attributes. In case of textual attributes collocation matrix row is used. In case of citation attributes co-citation matrix row is used.

For clustering VOS solves the same task [9], meaning same drawbacks. We propose to run metric clustering algorithms on resulting vectors.

Finally, to estimate the probability of co-authorship we propose to use Bayes rule. Let A be event that two authors have am article and B the event of occurrence of their meta-data together. We estimate prior distribution of A with:

$$\mathbb{P}\{A\} = \frac{\#\{\text{patents of this author}\}}{\#\{\text{patents total}\}} \tag{7}$$

Then, according to Bayes rule:

$$\mathbb{P}\{A \mid B\} = \frac{\mathbb{P}\{B \mid A\}\mathbb{P}\{A\}}{\sum\limits_{C} \mathbb{P}\{B \mid C\}\mathbb{P}\{C\}} \tag{8}$$

III. Evaluation

For demonstration purposes Internet of Things field was chosen. Patent data for research was acquired through European Patent Office (EPO) API. This tool allows to get data such as inventor names, assignee, different dates, classification, citations both ways, abstracts, claims and full texts. A corpus of 150 documents is used.

After vectorizing patent data in the way discussed earlier, we apply MDS using three proposed dissimilarity metrics and VOS-original association strength for comparison. All the images share the legend: blue dots account for US and WO region, red dots for CN region and green dots for KR region. The edges correspond to large number of common keywords (at least 0.1%).

Applying euclidean, cosine and TS-SS distance leads to graphs shown in Fig. 4, Fig. 5, Fig 6 respectively. All the graphs feature visible separation of US and CN regions.

In detail, euclidean distance unsurprisingly draws attention to extracted keywords as this subvector is the most dense. Therefore patents group when they share similar semantic profile, leading to topic distinction.

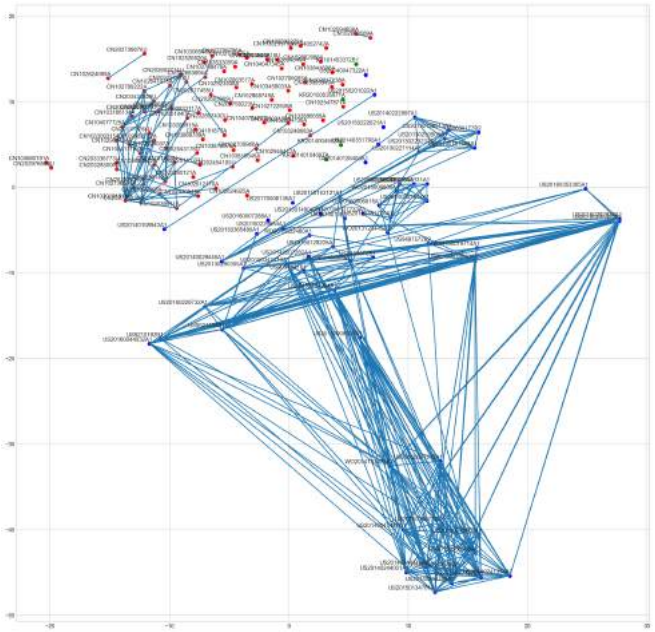

Figure 4. Euclidean distance.

Cosine distance, on the contrary, draws attention to classes, assignees and regions as they are represented as one-hot subvectors and affect angle rather than magnitude.

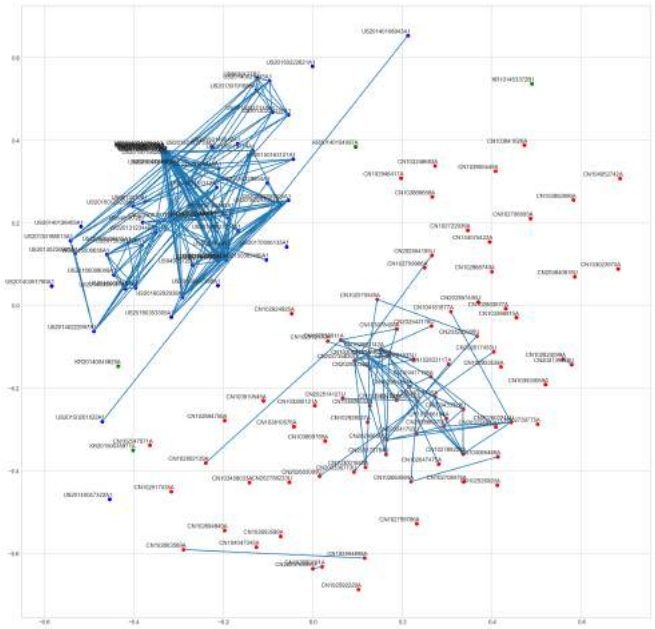

Figure 5. Cosine distance.

The TS-SS distance combines both features and is generally harder to interpret.

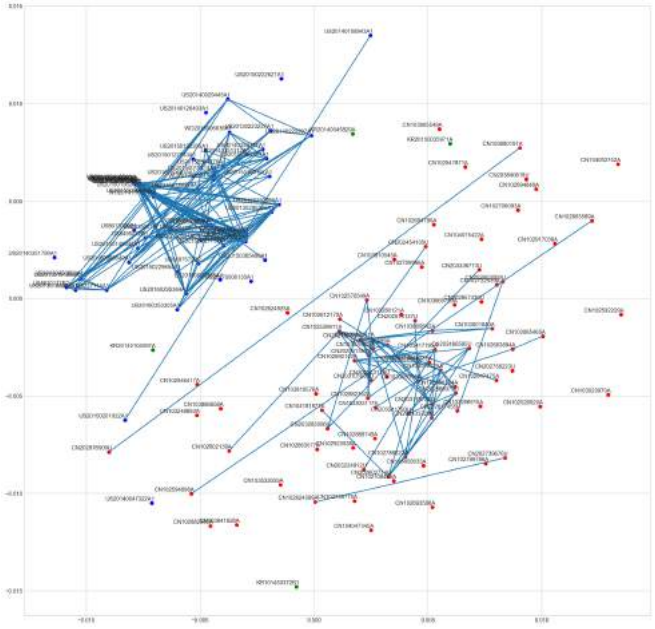

Figure 6. TS-SS distance.

IV. Further research

As an application of proposed model we see decision support systems in patent analysis. Russian GOST for patent analysis states the tasks, that should be included in this research, including:

- Research of a technical level of objects of economic activity, revealing of tendencies, a substantiation of the forecast of their development
- A study of the state of the markets for these products, the prevailing patent situation, the nature of national production in the countries of study
- A study of directions of research and production activity of organizations and firms that operate or can operate on the market of products under study
- Justification of proposals on the feasibility of developing new industrial property for use in facilities that ensure the achievement of technical indicators foreseen in the technical task (tactical and technical task)

Most of them are directly linked to analysis and forecasting of patent landscapes, which is successfully achieved with the help of proposed system. To even improve the system and minimize expert involvement, Deep Learning can be used to mine relations as in [10]. Some improvements may also be achieved with the help of latent semantics and topic modelling, as existing patent classification was only briefly introduced to proposed model.

V. Conclusion

In comparison to current research, the proposed method includes more complex and detailed intellectual analysis of patents, including implicit linguistic factors.

Acknowledgment

Author would like to thank Vladimir Khoroshevsky for problem statement and useful discussion.

References

[1] D. Bonino, A. Ciaramella, and F. Corno, "Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics," *World Patent Information*, vol. 32, no. 1, pp. 30 – 38, 2010.

[2] I. V. Efimenko, V. F. Khoroshevsky, and E. C. M. Noyons, *Anticipating Future Pathways of Science, Technologies, and Innovations: (Map of Science)2 Approach*. Cham: Springer International Publishing, 2016, pp. 71–96.

[3] R. Winiarczyk, P. Gawron, J. Miszczak, L. Pawela, and Z. Puchala, "Analysis of patent activity in the field of quantum information processing," vol. 11, 12 2012.

[4] I. Efimenko and V. Khoroshevsky, "Peaks, slopes, canyons and plateaus: Identifying technology trends throughout the life cycle," vol. 14, p. 1740012, 11 2016.

[5] V. F. Khoroshevsky and I. V. Efimenko, "Russia among the world centers of excellence," in *Open Semantic Technologies for Intelligent Systems (OSTIS-2016)*. BSUIR, February 2016, pp. 223 – 232.

[6] N. J. van Eck and L. Waltman, *VOS: A New Method for Visualizing Similarities Between Objects*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 299–306.

[7] N. J. van Eck, L. Waltman, R. Dekker, and J. van den Berg, "A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS," *CoRR*, vol. abs/1003.2551, 2010.

[8] A. Heidarian and M. J. Dinneen, "A hybrid geometric approach for measuring similarity level among documents and document clustering," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, March 2016, pp. 142–151.

[9] L. Waltman, N. J. van Eck, and E. C. M. Noyons, "A unified approach to mapping and clustering of bibliometric networks," *CoRR*, vol. abs/1006.1032, 2010.

[10] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 926–934.

ИЗВЛЕЧЕНИЕ И ВИЗУАЛИЗАЦИЯ ПАТТЕРНОВ ДАННЫХ ДЛЯ ПОСТРОЕНИЯ ПАТЕНТНЫХ ЛАНДШАФТОВ И ВЫЯВЛЕНИЯ НОВЫХ ТЕХНОЛОГИЧЕСКИХ ТРЕНДОВ В ОБЛАСТИ "ИНТЕРНЕТ ВЕЩЕЙ"

Николай Стулов
ФУПМ МФТИ

Извлечение сложных отношений из неструктурированных данных — это сложная задача в любой области, а в особенности в быстрорастущих областях, таких как Интернет вещей. В этой работе исследуются различные методы извлечения и визуализации этих отношений. В результате предлагается набор инструментов для обработки текстов и патентов, а также подход, который может быть использован для построения интеллектуальной системы поддержки принятия решений.