DOI: 10.2507/28th.daaam.proceedings.060

EXTRACTING PATENT-RELATED INFORMATION FROM ONLINE SOCIAL NETWORKS: CASE OF FACEBOOK

Alexander Ivanov[1,2] & Zeljko Tekic[1]

[1]Skolkovo Institute of Science and Technology, Moscow, Russia [2]National Research University Higher School of Economics, Moscow, Russia

This Publication has to be referred as: Ivanov, A[lexander] & Tekic, Z[eljko] (2017). Extracting Patent-Related Information from Online Social Networks: Case of Facebook, Proceedings of the 28th DAAAM International Symposium, pp.0433-0438, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-11-2, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/28th.daaam.proceedings.060

Abstract

In this study, we investigate how to identify and extract patent-related information from Facebook, the largest online social network. In the first step, we identified a list of trustworthy sources we started search from. Then, we developed algorithms for extracting and filtering information, and based on them software tool that is able to identify and deliver a patent, the Facebook post where it is mentioned and news / blog article which discuss it. We did a pilot test and collected more than 50 examples of articles (and Facebook posts) that add value to the patent they refer to. We classified collected articles and discussed how they can be used. Finally, we outlined where developed tool can be applied.

Keywords: Facebook; social networks; patent analysis; patents

1. Introduction

Since the beginning of this century social media and, especially online social networks (OSN) have exploded as platforms where users generate and share content and intensively engage with it through different actions. Online social networking sites like Facebook, Instagram, Twitter and LinkedIn attract more and more users every day. Being widespread, easy to use and available everywhere online social networks provide fast and powerful communication platform which sets trends and shapes public opinions in topics that range from politics and economy to technology and entertainment across societies.

OSNs have millions of members / users and these people possess different knowledge, expertise and skills sets, thus OSNs can be seen as a form of collective wisdom platform [1]. Bearing this in mind, we decided to investigate patents as a topic in OSNs. More specifically, we are interested in extracting patent-related information we are able to identify at Facebook, the largest OSN.

Why are we interested in patent-related information from Facebook? Patents are a powerful and unique source of data for innovation and technology analyses. However, extracting useful information from patents and freely available patent

databases is not easy. Identification of relevant and valuable patents is still difficult, time- and manpower-consuming work, which requires special expertise [2]. Current research and developed software tools in the field of patent analytics [3, 4] try to respond to this challenge using exclusively patent data (i.e., bibliographic data, patent descriptions, abstracts, claims, etc.). This approach is limited in many ways – it does not solve the problem of decoding and interpreting patent language, it rarely allows the matching of patents with specific product features and/or specific products, and it requires strong expertise in the field of intellectual property law. To improve practical value of available patent information and democratize its usage ("patents for people, not only for experts"), we propose to use contextual information, information that is related to specific patent but is not part of patent (document) itself [5]. Instead of limiting ourselves to using only patent data, we will combine patent data with relevant information from context in which specific patent is mentioned by people intrinsically or professionally interested in the topic. Our idea is to identify and extract value from posts in which people with interest in patents and patented technology write about a certain patents, their features, potential application areas and importance. This information provide context needed to easier understand patent, its language and, possibly, its value. This source of information is unexplored so far in patent analytics, and we believe that it has potential to bring significant value to the field. However, before we start analysing contextual information as an input in patent analytics, we need to collect it. The objective of the paper is to describe the structure and functions of recently developed software tool that is able to identify and extract patent-related information from Facebook.

The remainder of the paper is organized as follows. In Section 2 we review relevant literature, while in Section 3 we describe our approach and algorithms used. Section 4 presents results and discusses implications for practice and future applications. Finally, in Section 5 we conclude with a summary of results, limitations and the future research directions.

2. Literature Review

OSNs, with billions of users, represent a very interesting source of data. It is recognised by many researchers who used them in different ways. We can differ them by the sources of data which is used and the application.

The first direction is the information distribution, where the speed of analysis matters, for example, in stock analysis. Such a direction includes Twitter analysis [6] and Facebook analysis [7]. The second one analysed millions of Facebook pages in order to predict stock behaviour. Correlations between the sentiment and volatility were found.

The next direction is sentiment analysis, when one would like to figure out people's attitude to a certain subject, for example, regarding mobile phone providers [8]. Another paper [9] studied the sentiment analysis in e-learning, and in the sentiment detection part they combined two approaches: machine learning approach and lexicon-based approach. The lexicon-based approach consisted of using a dictionary of keywords marked with the sentiment they represent and finding using the keyword search the sentiment of each phrase splitting it into tokens.

Social networks allow researchers to analyse text in order to find peculiarities. Such research may be restricted to a certain domain. For example, [10] studied the way how to alleviate the depressive symptoms. Instead of taking millions of pages and scanning large sets of data, they took 68 participants of the experiment and monitored only their texts. Apart from sentiment analysis, they also took into account additional features, e.g. the number of friends, number of comments, etc.

Finally, it is possible to analyse influencers [11]. Such methods represent social network as a graph with nodes and edges, and the task is to find the most influential nodes of the graph in terms of information distribution.

Next, we would like to discuss the methods and algorithms which are used in analysis. Some papers study not the specific applications of the Facebook data, but the algorithms and software for crawling data from Facebook. Rieder [12] introduces a Netvizz app which allows to grab data from specific Facebook groups and build friendship graphs. The software uses Facebook API and automatically downloads the data specified by the user. Several examples of applications are provided.

Additional line of research is studying the efficiency of different crawling algorithms. For example, Ye et al. [13] studied the efficiency of four search techniques (BFS, greedy, lottery and hypothetical greedy) applied to four different sources of data (Flickr, LiveJournal, Orkut and Youtube).

The next part of the literature review is the patent analysis. Patents are helpful for strategic planning purposes [14]. The main problem is that hundreds of thousand patents are published every year, and the task is to select the most valuable ones is hard and time-consuming. The most popular approach is to analyse patent parameters and the ways they indicate the patent value. The most influencing parameter is the number of forward patent citations [15] showing the industrial importance of the patent. Lee and Sohn [16] study the first patent citations as a faster indicator than the number of all citations, because some citations are received 5-10 years after the patent publication.

In [17], Ivanov and Tekic, instead of using patent parameters, used data from blogs. Websites are searched for the articles about specific patents in order to obtain expert insights about patents. As a result, it is said that such articles might give a clearer understanding of a patent.

In research presented here, we study Facebook as a primary social network and as a source of data, because it has more text content than Twitter, and it is the largest social network in the world with more than two billion users. We use Facebook application programming interface (API) for downloading and searching for patent-related data, the process will be described in details in next sections.

3. Approach and Implementation

The workflow of our approach is represented at the scheme at Figure 1.

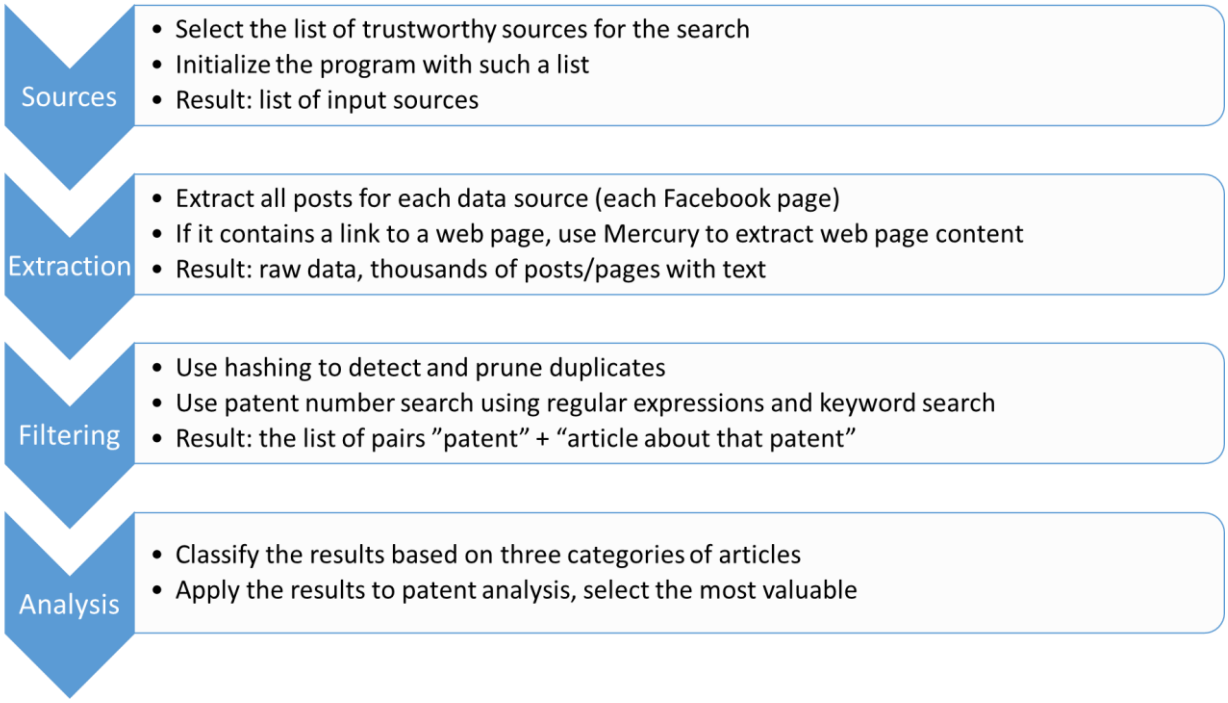

Fig. 1. The program workflow

Now we will discuss each of the steps from the workflow, one by one. Each data search must start from a list of entry points. It can be either the whole social network or a list of pages from a given social network. In the research connected to a specific topic, it might be a good idea to restrict the search domain. For example, in the patent research taking all Facebook pages may give a situation when not only expert opinions are taken into account, and the resulting dataset might include some irrelevant comments or even spam about patents. That is why it is better to take a list of trustworthy sources.

In the patent research we took only trustworthy sources by the following algorithms. First, we found several website catalogues about best websites devoted to intellectual property or technology. Second, for each website from the catalogue we looked for a Facebook page of this media. If it existed, we included that Facebook page into the initial list of search sources.

The next step is downloading data from social networks. Facebook provides a useful API [18] in order to access its data. Data which is available via API does not differ from the data one can see online on a certain page, but is represented in a structured way. The list of data which we extract includes the posts from a certain Facebook page, the number of likes, the number of shares and the number of comments. Facebook API might not be enough for the search in social networks. It turns out that many media projects do not include much information into their Facebook posts. In many cases a typical Facebook post of large and small media contains information about the title of the news and an external link to their website.

The next idea is to follow the link and to get information from the media website, but it is a complicated issue, because one will get an HTML web page which consists of not only text, but also the markup and a lot of sidebars and menus. Such information would be redundant for our research, we need only the content of the article which is mentioned in the Facebook post. For this reason we use additional tool for extracting content. Is it a library available for free use called Mercury [19]. For a given web page it returns the content of the webpage without sidebars, menus and most of the markup.

Next comes the text processing phase. It should select only relevant results from the whole set of found posts and pages. The phase consists of two steps: the keyword search step and removing duplicates step.

The keyword search step is aimed at finding the required keywords or word combinations in the text. In the case of patent search there are three situations when a certain piece of text is considered as a good result (by “good result” we mean an article about specific patent where it is possible to get the exact patent number the articles refers):

1. It contains a certain keyword (e.g. “patent”) and the patent number (e.g. “US 1,234,567”). This search is performed by a combination of keyword search and regular expressions search. We selected a list of possible representation of patent numbers in the text including small and capital letters, spaces or commas as separators etc. and developed a list of regular expressions. If a given article matches at least one regular expression from our library, the article is considered as a good one.

2. It contains a link to the patent database to a specific patent. In this case we need to find all external hyperlinks in the text and compare them to the list of patent databases. If at least one hyperlink in the text matches with the patent databases by domain name, we consider the article as a good result.

The keyword search step takes every found article as input and checks whether it contains a patent number or a link to the patent database. The result of the keyword search step is the set of potential good results. We would like to underline the point that the keyword search step is the step which defines the application of the search algorithm. The program is universal and may be used in searching not only patent related data, but also data about brands or political preferences. In this case, the keyword search step should be fine-tuned if we would like to change the course of the reseach, while all other steps may remain intact.

The removing duplicates step is aimed at pruning pages which copy the content of previously found pages. It includes the following possible situations:

1. The same page with the same text is stored by different URLs
2. A certain media decided to republish the same text on their website by simple copying the content without any new comments or text

The program should not include duplicate results, for this reason it should contain a special module which deals with such situations.

The brute-force solution of such problem is to compare every new found piece of text with all previously found useful articles. The problem of such approach is that symbol-per-symbol comparison for each found article will give a computational complexity of O(length(text)*number_of_found_good_articles) for each article which might contain useful information. If we assume that our program should find thousands of good articles where each article consists of thousands symbols, the resulting computational complexity may significantly hurt the speed of the algorithm.

In order to solve this problem we use hashing to increase the speed of search. After removing all punctuation marks and other symbols except for the letters we calculate the hash value of a certain article using the formula:

$$\sum_{i=1..n} s_i * c^{i-1} \; modulo \; p$$

where s_i is the i-th symbol of the text, c and p are two hashing constants. In order to decrease the number of possible collisions, we use double-hashing with $c = 31, p = 2^{32}$ and $c = 53, p = 2^{32}$.

How the two strings are compared using hashes? If both hashes are different (first hash for string1 is not equal to first hash of string2 and second hash of string1 is not equal to second hash of string2), we assume that the two strings are also different. It means that we store the hashes for all found good articles and for each new candidate we simply calculate the two hash functions and look for those values in the current results, determining whether we already found such an article or it is a new result.

4. Results Overview

The program for finding contextual data about patent has been running for two months during summer 2017. As a result, more than 300,000 Facebook posts from 108 pages were analysed and 53 posts about specific patents were found. The average speed of the program is one parsed Facebook page with all its posts per one day if it is launched on one thread on one computer. This includes running all algorithms for finding the patent numbers, removing duplicates, hashing, etc. The run-time can be decreased if the program is launched concurrently on two or more computers using two or more threads on each computer, so the overall processing time can easily be reduced to several days.

In this Section we will classify the results. The 53 found pairs “article” + ”patent” can divided into three groups:

1. Descriptive articles about patents. This category stands for articles which describe a recently issued patent (issued after 2004) and its possible applications.
2. Articles about historical / old patents. This category stands for patents which were published more than 13 years ago (before 2004). Why this threshold was chosen? The main reason is the scope of the research. We use Facebook as a source of acute data, Facebook was started in 2004, it means that before 2004 no media could publish a news article on Facebook about recently issued patent.
3. Articles about litigation. If Apple sues Samsung, the case is usually covered by a lot of websites and blogs and has many posts about it in social networks. If there is an article about a patent issued before 2004 and related to litigation, it will go to this category.

The most common case is the article about litigation. Approximately half (51%) of the results which we found are the articles about a litigation case with a patent involved. 21% of the results are articles about “old patents”, i.e. patents issued before 2004. In this case, we cannot track the speed of media reaction to the publication and the value of the patent or we will get the biased results, because our source of data simply did not exist when the patent was published. Finally, 28% of found articles are the descriptive articles about specific patents. This is the most valuable part of the results. Why? Because instead of reading a patent for two hours a person without specific knowledge in the field may get the understanding of the substance of the patent.

The classification made based on the 53 found articles is presented at Figure 2.

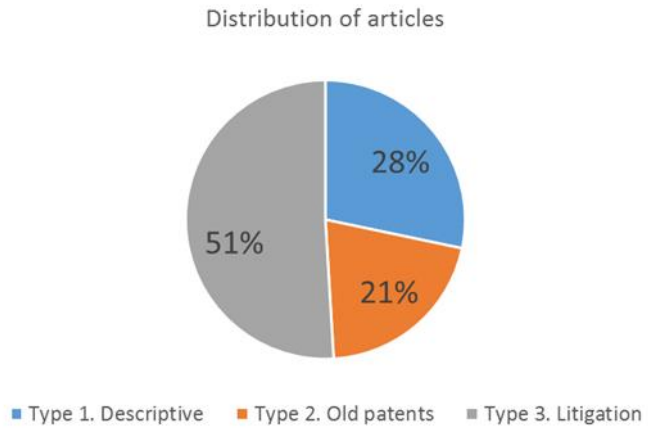

Fig. 2. Distribution of the found articles

Let us provide an example:

Patent number: US 8,253,639
Patent title: Wideband electromagnetic cloaking systems
Link to the patent: www.freepatentsonline.com/8253639.html
Link to the article: www.ipwatchdog.com/2012/09/07/uspto-issues-worlds-first-invisibility-cloak-patent/id=27841/

The article contains a comment of the inventor, a descriptive video and a prediction about its further development. Everything can be read within two minutes. Each out of 15 found articles of the category "descriptive articles about a specific patent" shows a more concise and easily readable description of a certain patent.

The use-case of such an approach may include a software which searches for experts' opinions for specific patents issued not so much time ago. Small and medium companies which do not have much budget for the intellectual property research may use it to monitor the competitive landscape. With thousands patents published every day, such a tool can decrease the amount of time needed for a person to understand each patent. With an easy understandable expert opinion it would take several minutes per patent instead of two hours.

5. Conclusion

Social networks represent a source of data which can be applied to different types of research. In this study, we investigated how to identify and extract patent-related information from Facebook, the largest online social network. We developed algorithms for extracting and filtering information, and based on them software tool that is able to identify and deliver a patent, the Facebook post where it is mentioned and news / blog article which discuss it. We did a pilot test and collected more than 50 examples of articles (and Facebook posts) that add value to the patent they refer to. We classified collected articles and discussed how they can be used.

The next steps of the research can be enhancing the list of data sources with more Facebook pages as well as adding websites to the list of sources. Additionally, the engagement numbers, e.g. the number of likes and comments should be studied as potential indicators of patent value.

The designed software tool is universal and may be applied not only to patent search, but also to all other types of keyword or regular expressions search. For example, the designed software can be easily switched to a brand-awareness search, when the task is to analyse how the crowd assesses the products of a certain brand.

What are the limitations of the approach? First, data from social networks represent data generated by all users of that social network. It means that if we need precision and accuracy, we need to restrict the list of sources or to check each source of information. In our case, we selected the ranking of media and took their pages on Facebook.

Second, the data on social networks related to specific topic is rather sparse. After processing 300,000 posts from the selected list of media, we found only 53 articles about specific patents. For example, if one would like to create a keyword monitor like a brand-loyalty monitor in social networks, he will need a lot of computational resources to get a representational set of opinions about his brand.

Third, most media put a link to their website and include only the title of the publication in the Facebook posts. It means that a direct web search, when not the pages on Facebook are analysed, but the media websites themselves, might give the same or even more results. In terms of applying contextual data analysis to patent intelligence, the next step might be the data search from the whole Internet or from the selected list of websites in order to find data about specific patents. Such an approach may give more than 53 received in this research results.

6. References

[1] S. Asur and B. A. Huberman, “Predicting the Future with Social Media,” 2010 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., pp. 492–499, 2010.

[2] Z. Tekic and D. Kukolj, “Threat of Litigation and Patent Value: What Technology Managers Should Know,” Res. Manag., vol. 56, no. 2, pp. 18–25, 2013.

[3] A. Abbas, L. Zhang, and S. U. Khan, "A Literature Review on the State-of-the-art in Patent Analysis." World Pat. Inf., vol 37, pp. 3–13, 2014.

[4] Z. Tekic, M. Drazic, D. Kukolj, and M. Vitas, “From Patent Data to Business Intelligence – PSALM Case Studies.”. Procedia Eng. vol. 69, pp. 296 – 303, 2014

[5] A. Ivanov, and Z. Tekic, “Towards smart patent analytics - matching patent and contextual data”, paper presented at R&D Management Conference 2016 “From Science to Society: Innovation and Value Creation” 3-6 July 2016, Cambridge, UK

[6] F. Corea, “Big Data Research Can Twitter Proxy the Investors ’ Sentiment ? The Case for the Technology Sector,” Big Data Res., vol. 1, pp. 1–5, 2016.

[7] A. Siganos, E. Vagenas-Nanos, and P. Verwijmeren, “Facebook’s daily sentiment and international stock markets,” J. Econ. Behav. Organ., 2014.

[8] N. A. Vidya, M. I. Fanany, and I. Budi, “Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers,” Procedia Comput. Sci., vol. 72, pp. 519–526, 2015.

[9] A. Ortigosa, J. M. Martín, and R. M. Carro, “Sentiment analysis in Facebook and its application to e-learning,” Comput. Human Behav., vol. 31, no. 1, pp. 527–541, 2014.

[10] S. W. Lee, I. Kim, J. Yoo, S. Park, B. Jeong, and M. Cha, “Insights from an expressive writing intervention on Facebook to help alleviate depressive symptoms,” Comput. Human Behav., vol. 62, pp. 613–619, 2016.

[11] E. Lahuerta-Otero and R. Cordero-Gutierrez, “Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter,” Comput. Human Behav., vol. 64, pp. 575–583, 2016.

[12] B. Rieder, “Studying Facebook via data extraction,” Proc. 5th Annu. ACM Web Sci. Conf. - WebSci ’13, pp. 346–355, 2013.

[13] S. Ye, J. Lang, and F. Wu, “Crawling online social graphs,” Adv. Web Technol. Appl. - Proc. 12th Asia-Pacific Web Conf. APWeb 2010, no. February, pp. 236–242, 2010.

[14] H. Ernst, “Patent information for strategic technology management,” World Pat. Inf., vol. 25, no. 3, pp. 233–242, 2003.

[15] M. B. Albert, D. Avery, F. Narin, and P. McAllister, “Direct validation of citation counts as indicators of industrially important patents,” Res. Policy, vol. 20, no. 3, pp. 251–259, 1991.

[16] J. Lee and S. Y. Sohn, “What makes the first forward citation of a patent occur earlier?,” Scientometrics, pp. 1–20, 2017.

[17] A. Ivanov and Z. Tekic, “Using Contextual Data for Smart Patent Analysis,” 2016 IEEE Int. Conf. Cloud Comput. Technol. Sci., pp. 448–451, 2016.

[18] “Facebook API for Developers.” [Online]. Available: https://developers.facebook.com/. [Accessed: 26-Sep-2017].

[19] “Mercury Web Parser.” [Online]. Available: https://mercury.postlight.com/web-parser/. [Accessed: 26-Sep-2017].