HIM-PRS: A Patent Recommendation System Based on Hierarchical Index-Based MapReduce Framework

Xuhua Rui and Dugki Min(✉)

Department of Computer, Information and Communication Engineering,
Konkuk University, Seoul, South Korea
{abealasd, dkmin}@konkuk.ac.kr

Abstract. Intellectual Property (IP) data, such as patent documents, grows inconceivably in recent years. Therefore, discovering valuable information from those huge number of data becomes a challenge. This paper introduces a novel patent recommendation system called HIM-PRS which is built on top of hierarchical index based big data processing platform. HIM-PRS integrates with linked data to provide an efficient patent recommendation service. Our result shows that HIM-PRS is able to find more semantically similar patents than other systems. Additionally, HIM-PRS launches query jobs at least 2 times faster than original Hadoop MapReduce framework.

Keywords: Patent recommendation service · HIM · BigData Processing · Linked data

1 Introduction

With rapid development of technologies, Intellectual Property (IP) plays a more and more important role in our life. Because it represents the legal power for people to protect their invisible treasure. Last year, more than 7500 patent disputes happened in the U.S., and more than half of them are related with high technology [1]. However, an intellectual property dispute will cost a various of long term lawsuits. For example, Apple Inc. started to sue Samsung Electronics Co., Ltd. in the year of 2010. Finally, various of lawsuits in more than 9 countries are involved in their IP war during the past 5 years [2].

There are a lot of reasons making those disputes so complicated, durable and difficult. The most important reason is the rapidly increased number of patents cannot be analyzed within a short time. In the year of 2000, only 175979 patents have been granted by United States Patent and Trademark Office (USPTO). 4 years ago, the number became 276788. During the past decade, more than 3 million patents has been officially granted by USPTO [3].

Discovering meanings from those amount of data challenges more and more IT companies. USPTO has opened all U.S. patent documents to public and provided a lot of methods to access their database [4]. However, the query performance limits users. From 2010, Google, the global internet information searching service provider,

J.J. (Jong Hyuk) Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 421, DOI 10.1007/978-981-10-3023-9_130

co-worked the USPTO to provide their bulk data for public access [5]. Google also opened a patent searching service [6] for helping users to discovery IP information easily.

Since such amount of patent document data has been opened, researchers apply various of approaches to discovery treasures from them. Statistical patent analysis approach [7] has been widely used for simply measuring and visualizing patents based on technology catalog. Citation based patent analysis [8] approach is inspired by information retrieval technology. It is able to present inner-relations of patents in addition. Patent cluster analysis is another patent analysis approach. By utilizing cluster analysis approach, patent documents can be organized as a theme map which describes patents relationship simpler [9]. Furthermore, researchers are interested in utilizing ontology technology to discovery knowledge from patent data [10].

However, there is still not a sufficient solution for analyzing patent big data based on the semantic meaning of patents. In this paper, we will introduce our proposed system, HIM-PRS, which is designed for analyzing patent bigdata semantically.

The rest of the paper is organized as follows: We first show the architecture of HIM-PRS, and afterward, we describe the system working mechanism in Sect. 2. In Sect. 3 we present the experimental results of our proposed system with analysis. Finally, we give the conclusion.

2 HIM-PRS Architecture

In this section, we mainly focus on introducing the design of our proposed patent recommendation system: HIM-PRS.

2.1 HIM: Hierarchical Index-Based MapReduce Framework

The fundamental of our proposed system is a bigdata storage and processing system which is called HIM. HIM is inspired by Google's MapReduce paper [11] and built on top of the most widely used open source bigdata solution Hadoop [12].

The original Hadoop keeps data uploaded by users in separated blocks. However, as time flies, people realized that the plain data storage structure is no longer sufficient for big document analysis. Database technologies performance a good result for data retrieval by using index. However, as the data grows, the index generation time and size become unacceptably. Therefore, many researchers try to integrate the advantages of both technologies, such as HadoopDB [13], HBase [14], Hive [15], Hadoop ++ [16] and BigTable [17]. Inspired by previous researchers' work, we proposed a hierarchical index-based patent data storing mechanism [18].

Based on the previous proposed mechanism, we built our hierarchical index-based MapReduce framework for storing and processing patent documents.

Figure 1 shows the main concept of HIM. Since HIM is built on top of Hadoop, there are 2 kind of servers are involved. A server works as a master node which maintains the entire distributed system(DS) information. A master node is in charge of creating, updating, deleting, backup and recovering the DFS Hyper Index. DFS Hyper

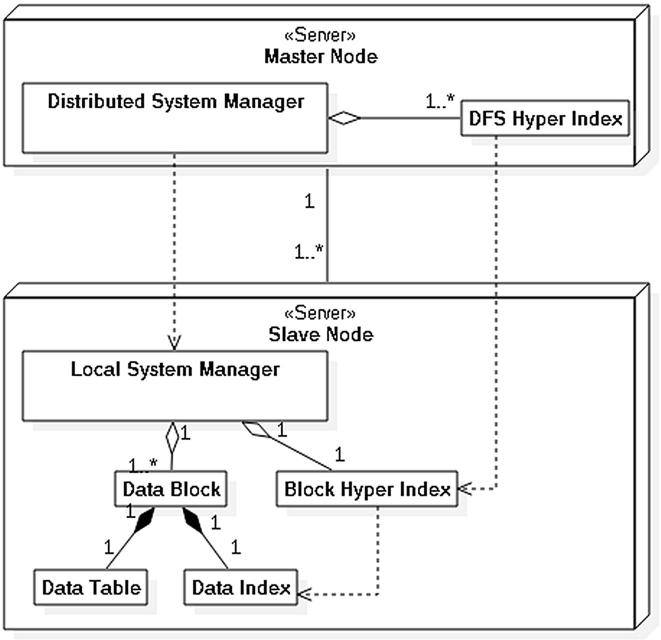

Fig. 1. Concept view of hierarchical indexes in HIM

Index is a high level index which keeps the hot data index collected from Block Hyper Index. Similar like Hadoop, real data are still stored in slave nodes organized in data blocks. Data blocks in HIM organize original data in table structure and generate index for each data table additionally. As a result, data blocks in HIM are dynamically changed based on user requirement. Each slave node also maintains a block hyper index which contains all hot data index collected from data index inner data blocks.

Integrating the hierarchical index based data storing mechanism with Hadoop's MapReduce data processing framework, our proposed HIM is able to do data analysis works more efficiently than the original Hadoop.

2.2 HIM-PRS: HIM Based Patent Recommendation System

On top of HIM, we built a Patent recommendation system which is designed for high speed patent searching and semantic patent recommendation. Instead of using 2 type of servers, HIM-PRS requires an additional control server for handling the entire patent analysis and service processing flow.

Our proposed system is designed to interact with a global scale knowledge base provided by linked data project. Therefore, SPARQL is the default knowledge query language used by HIM-PRS for communicating with remote knowledge node.

As Fig. 2 shows the HIM-PRS architecture. HIM-PRS control node maintains all modules related with our patent services. To provide our service, firstly, Patent Feature

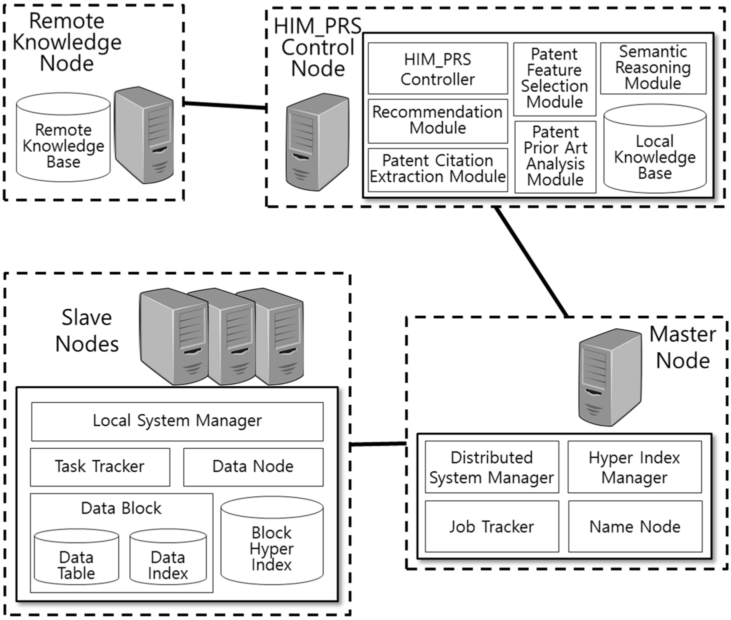

Fig. 2. HIM-PRS architecture

Select Module and Patent Citation Extraction Module analyze patents and extract necessary information such as patent feature words and citation data. Then, Patent Prior Analysis Module will generate patent prior art graph based on patent citation information. These 3 modules are pre-processing modules. Secondly, Recommendation Module receives keywords from users and transfers those keywords to Semantic Reasoning Module. The Semantic Reasoning Module first queries local knowledge base via SPARQL, and if there is no result found, it directly requests to remote knowledge base via the same SPARQL query.

3 Experimental Results and Analysis

In this section we will show a performance test of HIM and a patent recommendation example. We utilize total 14 intel dual core machines for running our experiment. Machines equipped with 2 GB memory work as HIM-PRS Control Node and HIM Master Node respectively. And the rest machines work as slave nodes. All machines are connected through a 100Mbit switch. Approximate 106 GB patent data which includes 1328892 patent documents collected from Google bulk data have been uploaded to our HIM-PRS before our experiments.

Figure 3 shows the query time of different block size. The blue line represents query time of using non-indexed data block while the red line represents using indexed date block. Obviously, as the data size growth, the query time for data in non-indexed blocks increases linearly. In the other side, using indexed blocks the query time

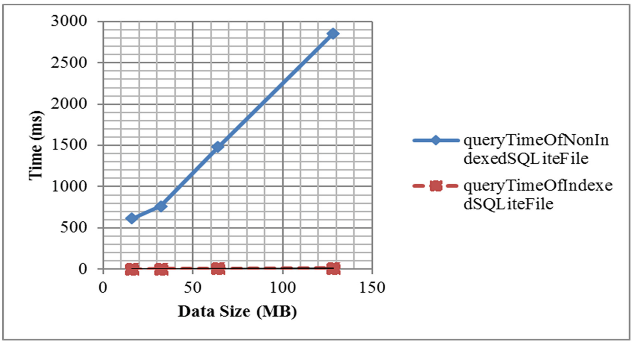

Fig. 3. Data query time for different block size

increases slightly. The non-indexed approach represents the original Hadoop solution. Our proposed HIM approach performs a better data query performance than the original Hadoop approach. Considering time for job preparation, jobs launching on top of HIM perform at least 2 times faster than original Hadoop. On top of HIM testbed, we built a HIM-PRS prototype. Currently we utilize WordNet [19] as our local knowledge base and DBpedia [20] as our remote knowledge base. In our example, we try “cloud computing” as our input key word.

Our proposed system utilizes a SPARQL query to query word “cloud” from local knowledge base WordNet. Similarly, we query the word “computing” also. Combined results of previous queries, we get a group of external keywords. Then, we send a query about “cloud computer” to remote knowledge base which is provided by DBpedia project through SPARQL. We get another group of semantic related words. By combining both groups of words, we have our final keywords. Finally, we searching patents which contain those keywords and ranking them based on our prior art analysis result. Based on system patents 7225249 and patent 8341462. The first patent claims a web based application which works under a PaaS pattern and the second patent claims a cloud provisioning technology. Both them are related with cloud computing topic and none of them shown in google patent searching page.

4 Conclusion

In this paper, we present a patent recommendation service system which is built on top of a performance enhanced Hadoop platform: HIM. Our proposed system performs a better patent data query performance. Meanwhile, our proposed system considers the semantic meaning of input keywords for patent recommendation.

We compare our platform with original Hadoop platform and google patent searching service. Our system shows a better data query performance than original Hadoop. However, if the query word is not indexed, our query will still as same as original Hadoop. In fact, if we build index for all data, the index size will be incredible. Thus to balance the size of index and performance of querying data is a new challenge for index based MapReduce platform and which is our future work.

Acknowledgement. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the University Information Technology Research Center support program (IITP-2016-R2720-16-0004) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Patent Dispute Report (2015). http://www.unifiedpatents.com/news/2016/5/30/2015-patent-dispute-report
2. Apple Inc. v. Samsung Electronics Co. https://en.wikipedia.org/wiki/Apple_Inc._v._Samsung_Electronics_Co
3. USPTO. U.S. Patent Statistics Chart: Calendar Years 1963–2015. http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm
4. USPTO. USPTO Patent Search, http://www.uspto.gov/patents/process/search/
5. Google. USPTO Bulk Downloads: Patents. https://www.google.com/googlebooks/uspto-patents.html
6. Google. Google Patent Search. https://www.google.co.kr/#tbm=pts&gws_rd=cr
7. Mancusi, M.L.: Technological specialization in industrial countries: patterns and dynamics. Weltwirtschaftliches Archiv. **137**(4), 593–621 (2001)
8. Narin, F., Carpenter, M.P., Woolf, P.: Technological performance assessments based on patents and patent citations. IEEE Trans. Eng. Manage. **1984**(4), 172–183 (1984)
9. Aurek. http://starwars.wikia.com/wiki/Aurek
10. Taduri, S., et al.: A patent system ontology for facilitating retrieval of patent related information. In: Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance. ACM (2012)
11. Dean, J., Ghemawat, S.: MapReduce: a flexible data processing tool. Commun. ACM **53**(1), 72–77 (2010)
12. White, T.: Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol (2012)
13. Abouzeid, A., et al.: HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proc. VLDB Endow. **2**(1), 922–933 (2009)
14. George, L.: HBase: The Definitive Guide. O'Reilly Media Inc., Sebastopol (2011)
15. Thusoo, A., et al.: Hive: a warehousing solution over a map-reduce framework. Proc. VLDB Endow. **2**(2), 1626–1629 (2009)
16. Dittrich, J., et al.: Hadoop ++: making a yellow elephant run like a cheetah (without it even noticing). Proc. VLDB Endow. **3**(1–2), 515–529 (2010)
17. Chang, F., et al.: Bigtable: a distributed storage system for structured data. ACM Trans. Comput. Syst. (TOCS) **26**(2), 4 (2008)
18. Rui, X., Kim, B., Min, D.: An efficient patent storing mechanism based on SQLite on Hadoop platform. In: Chen, Y., Balke, W.-T., Xu, J., Xu, W., Jin, P., Lin, X., Tang, T., Hwang, E. (eds.) WAIM 2014. LNCS, vol. 8597, pp. 382–392. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11538-2_35
19. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
20. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). doi:10.1007/978-3-540-76298-0_52