UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

Relatórios Técnicos
do Departamento de Informática Aplicada
da UNIRIO
n° 0001/2022

A Systematic Mapping Study on Search Engine Industry's Patent Innovations Focused on Matching the Search Results to User Intent

Marcelo Tibau
Sean Wolfgand Matsui Siqueira
Maria Augusta S. N. Nunes
Bernardo Pereira Nunes

Departamento de Informática Aplicada

UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
Av. Pasteur, 458, Urca - CEP 22290-240
RIO DE JANEIRO – BRASIL

Relatórios Técnicos do DIA/UNIRIO, No. 0001/2022 January 2022

A Systematic Mapping Study on Search Engine Industry's Patent Innovations Focused on Matching the Search Results to User Intent

Marcelo Tibau[1] Sean W. M. Siqueira[1]
Maria Augusta S. N. Nunes[1] Bernardo Pereira Nunes[2]

[1] Depto de Informática Aplicada – Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
[2]College of Engineering & Computer Science, Australian National University, Australia

marcelo.tibau@uniriotec.br, sean@uniriotec.br, gutanunes@uniriotec.br,
Bernardo.Nunes@anu.edu.au

Abstract. This report presents a systematic mapping study to survey patent documents in Brazilian and international databases to characterize the technical state-of-the-art in search engines under the intent matching paradigm as well as to discover the size of academia initiatives as part of the technical state-of-the-art agenda to commercial innovation ventures in the search engine industry. The goal is achieved by the identification of the search engine industry most common innovation claims concerning the intent matching paradigm and by the discussion of important issues regarding its patent application patterns, such as the search engine industry's proficiency to churn its innovative landscape.

Keywords: Search Engine; Patent; Systematic Mapping; Innovation; Intent Matching.

* This study was financed in part by the "National Council for Scientific and Technological Development (CNPq) - Brazil" - Process 315374/2018-7, Project "Searching as Learning: the information search as a tool for learning" and by the "Coordination for the Improvement of Higher Education Personnel" (CAPES) – Brazil – Finance Code 001.

Table of Contents

1 Introduction

Web search engines help internet users sort out irrelevant content by showing documents that most likely fit users' search queries. The original search engine's paradigm is based on string matching (i.e., matching the query terms to document content), and the search engines leave the hardest task of reading and summarizing the relevant hits to the person who searches. Changing search engine's paradigm to matching the search results to user's information needs (i.e., intent matching) rather than pure string matching is a continuous research trend to enable various search settings that better assist users' searching needs (Zhang et al. 2020). Understanding users' right intention is pivotal to provide more user satisfying results (Mohammadi et al. 2020), which allows the search engine development to focus on the mechanics of how content is crawled, indexed, ranked, served, and evaluated within this new perspective.

Query intent detection (Mohammadi et al. 2020), user intent analysis (Zhang et al. 2020), and deep neural networks for user intention identification (Khattak et al. 2020) are some of the instances researchers apply in academia to enable the paradigm change to intent matching. Nevertheless, is it reaching search engine users as implemented innovations? More importantly, have search engines' innovations reflected the paradigm change throughout the years? The best tool to identify a given technology' evolution is patent databases. According to the World Intellectual Property Organization (WIPO), more than 70% of everything that has been published in terms of technology is found in patent documents (World Intellectual Property Organization 2019). Patents contain technical information about inventions and are legal instruments that confer rights and privileges to their owners (Clarke 2018), then, its information can be commercially important (e.g., to check whether someone has come up with the same ideas before or to know if other similar inventions exist). It is also a good way to determine how an invention works and its details of products and processes (Clarke 2009, p.119).

There can be different perspectives regarding a proper definition of innovation (Dziallas and Blind 2019). We apply the term to a successfully commercialized new idea (Dziallas and Blind 2019) once patents are only granted if the invention is novel (new), inventive (not obvious) and industrially useful (Clarke 2009, p.117). In this context, patent searching can aid at the discovery of "innovation footprints", with the term specifically referring both to innovative ideas intended to be commercialized (patent applications or patent just granted) and ideas that have already been successfully commercialized (patents prior granted). Knowing that there is an apparent mismatch in Brazil regarding academia outputs and innovation outputs in different knowledge domains (Procaci et al. 2016), it would be interesting to compare "innovation footprints" in Brazil and abroad concerning search engine's new paradigm of intent matching. We want to understand how search engines' matching capability evolved technically through patents and how much of those patents were from academia. As there are different approaches to characterize a scientific or innovative landscape (e.g., systematic literature reviews or systematic mapping studies) and we would like to understand what has been addressed by the community in a commercial capacity, a mapping study was defined as the method for structuring this work.

This report aims to characterize the technical state-of-the-art of methods employed to score documents and search queries gathered in the mapping study and grade the search in search engines through patent documents. It also intends to discover the academic initiatives' size as part of the technical state-of-the-art agenda to commercial innovation ventures in the search engine industry.

Including this introductory first Section, this paper is organized as follows: Section 2 describes the research method used for the mapping study presented in this paper. Section 3 presents the mapping results; Section 4 discusses the main findings regarding this study and Section 5 presents the threats to validity. Finally, Section 6 summarizes this paper's main contribution and discusses some research directions that could be investigated based on the patent application trends.

2 Methodology

This study applies the Systematic Mapping Studies[1] (Petersen et al. 2008) method to manage the process activities. The guidelines used were based on (Silva et al. 2018a) and (Silva et al. 2018b). A systematic literature mapping study' main objective is to provide an overview of a research area. In general, a systematic mapping study intends to structure the type of research reports and results that have been published by categorizing them and yielding a visual summary of its results (i.e., the map) (Wortmann et al. 2017). This method's essence is the definition of research questions, the conduct of search for relevant information (in our case, patent documents), the screening of documents based on inclusion and exclusion criteria, and the selection of relevant documents; followed by evaluation of their quality.

Patent documents were considered to identify technical state-of-the-art, and the applied patents survey was carried out in two different patent databases, WIPO's[2] Patentscope and INPI's[3] patent database. The online platform Parsifal[4] was used to organize the main activities and plan the protocol. The mapping scope was defined using PICOC[5] (Wohlin et al. 2012, p.45). Finally, to analyze the data and plot charts related to the results, we used Python scripts. The following subsections describe the methodology in detail.

[1] "Study" is the term employed by the Systematic Mapping Studies method to identify the set of literary studies in each area, whether they are state-of-the-art papers, technical state-of-the-art patents or other literary artifacts defined by the author.
[2] WIPO (World Intellectual Property Organization) is the world's number one source for global intellectual property information, resources, and services (www.wipo.int).
[3] INPI (Instituto Nacional da Propriedade Industrial - National Institute of Industrial Property) is the Brazilian office for industrial property that controls the issue of patents (https://www.gov.br/inpi/).
[4] https://parsif.al/
[5] PICOC is a method used to describe the five elements of a searchable question, and the name is an acronym that stands for population, intervention, comparison, outcomes, and context.

2.1 Research Questions

To provide an overview of the innovations regarding the search engine industry and to examine if the methods employed demonstrate the paradigm change from pure string matching to intent matching, we defined as a Primary Research Question: "Do search engine industry's innovations reflect the paradigm change detected in academia outputs?". We also considered a series of Secondary Research Questions to identify the quantity and type of patent application available within the retrieved results. Both Research Question types have the purpose of uncovering information presented in the analyzed patent application documents that respond as directly as possible to the study's goal. The defined questions and data are shown in Table 1.

2.2 The Search String and Databases

We designed a search string based on the PICOC (Wohlin et al. 2012, p.45) framework to increase the possibility of finding relevant patent application documents to answer the research questions. PICOC aided in defining this study's scope (shown in Table 2). As our focus of interest is the Search Engine industry, we chose it as the study's population. We defined the intervention (score document and grade the search) and the outcome (method, algorithm, computing, systems, and apparatus) clauses based on the machinery related dimension (Buganza and Della Valle 2010, p.45), which is one of the three types[6] of dimensions used to evaluate the search engine industry's performance. The study's context was defined as information retrieval, as it is the computer science subdomain in which search engine is included. As the comparison step (to which the intervention is compared) is not applicable, it was left blank.

Based on the terms from Table 2, we arranged a search string using the Boolean operators OR/AND (Table 3). In essence, this search string is a conjunction of three disjunctions. The first part of the conjunction specifies the study's subject of interest. The second part shows the features under study. Finally, the third part details terms representing the objects to be reviewed by the study.

We chose two databases (INPI's patent database and WIPO's Patentscope) to enable an overview concerning the technical state-of-the-art about search engine innovations in Brazil and abroad. INPI, being the office responsible for Industrial Property rights in Brazil, is a natural choice to search patents applied or granted in Brazil. As patent documents are filed in INPI only in the Portuguese language and its database concerns only patents applied in Brazil, the search was also conducted in WIPO's Patentscope database. Patentscope's choice is based on the necessity to search for a set of patent documents spread over a significant number of countries. According to their website[7], 91 million patent documents can be searched. To get a far-reaching outcome of documents to explore, we searched for patent documents in English and Portuguese lan-

[6] The three types of dimensions are (1) user-based, meaning what the user can perceive and act upon; (2) machinery related, meaning what the search engine does internally; and (3) business model-oriented,meaning what makes the business profitable.
[7] https://patentscope.wipo.int/search/en/search.jsf

guages and full-text format. As the Portuguese version of the search string returned no results matched, we decided to use only the English string.

We applied the search string in Patentscope, as shown in Table 3, with adaptations to fit INPI's patent database particularities. As its search engine did not authorize long Boolean search strings, we divided the search string into five little strings to get the original string effect from Table 3. Table 4 shows the search strings in Portuguese and their English version. It is important to point out that only the strings in Portuguese were applied.

2.3 Evaluation Criteria (Inclusion and Exclusion)

The inclusion of a patent document into the classification phase of this systematic mapping study was decided based on the reading of its title and abstract. We defined the following inclusion criteria (IC) and exclusion criteria (EC) to reduce the corpus and enable the study reproducibility. The inclusion criteria were grounded on the evaluation criteria proposed by (Buganza and Della Valle 2010, p.54-61) to the search engine industry's machinery related dimension:

2.3.1 Inclusion Criteria (IC)

- ***IC1:*** The patent application regards a pre-filtering step innovation, which means that it provides users with pre-filtering tools that give them the possibility to get to the desired results without entering any word or with little effort (e.g., query suggestion, auto-completion).
- ***IC2:*** The patent application regards a request step innovation, which means that it concerns how the user makes clear what she is looking for concerning the type of input allowed in the search engine (e.g., text, image, audio).
- ***IC3:*** The patent application regards a request processing step (i.e., query processing) innovation, which means technical activities that are put in place before executing the actual search (e.g., stemming, attempts to detect user intention).
- ***IC4:*** The patent application regards a data preparation step innovation, which means technical activities related to crawling, analyzing, indexing, and ranking resources.
- ***IC5:*** The patent application regards a search step innovation, which means technical activities matching the query against the indexes constructed by crawling Webpages.
- ***IC6:*** The patent application regards a result presentation step innovation, which means technical activities related to the presentation of the retrieved results (e.g., the Search Engine Results Page - SERP - features).
- ***IC7:*** The patent application regards a navigation step innovation, which encompasses actions for refinement and manipulation of output data (e.g., classification filters).

2.3.2 Exclusion Criteria (EC)

- ***EC1:*** The patent application document is not about a search engine industry innovation.
- ***EC2:*** The patent application does not provide enough information about its claim.
- ***EC3:*** The patent application document is not written in English or Portuguese or does not provide an English translation of its abstract.
- ***EC4:*** The patent application document does not present an innovation that could be identified as part of the machinery-related dimension of the search engine industry.

Hence, we analyzed the results considered suitable with further reading of the patent application document's claim and specification sections based on the inclusion criteria (IC) and exclusion criteria (EC). We then applied a quality assessment checklist with five quality questions (QQ). The cited checklist is presented in Table 5, with the possible score available for each question. The Quality Assessment Score results from the sum of the scores for each question, with a maximum score of 5.0 points. We also established a rejection score (2.0), in which a document would be rejected if left with a score less than or equal to it (a situation that did not happen).

The searches were performed in WIPO's Patentscope on September 18, 2020, and in INPI's database on September 24, 2020. An update search was performed on January 20, 2022, in both databases. A patent document was considered repeated when it possessed the same ID number of another retrieved document. Table 6 shows the systematic mapping search outcome and the ensuing evaluation.

3 Results

From the search strings shown in Tables 3 and 4, we were able to retrieve a total of 713 patent application documents from both databases. A total of 121 results were deemed repeated, which allowed us to analyze 592 single results (570 from WIPO's Patentscope and 22 from INPI's patent database) by reading their title and abstract. A total of 75 results were considered within the study's scope based on the inclusion criteria (IC), exclusion criteria (EC), and Quality Questions (QQ). The 64 patent application documents from WIPO's Patentscope and the 11 patent application documents from INPI's patent database are listed in the Appendix. Table 7 shows the selected patent documents from each database. A code identifies the database, WPS for WIPO's Patentscope and IPD for INPI's patent database, followed by the patent document's ID number in round brackets.

In the current Section, the systematic review results from the Secondary Research Questions are presented. The Primary Research Question will be discussed in Section 4.

3.1 What are the inventions that the search engine industry most frequently claims as innovations in patent documents (SRQ1)?

To provide an overview, we decided to organize the data into ten categories of innovation type derived from the following evaluation steps presented at the search engine industry's machinery related dimension (Buganza and Della Valle 2010): data preparation (p. 58), search (p. 59), and navigation (p. 60-61). Figures 1 and 2 show the categories of each database.

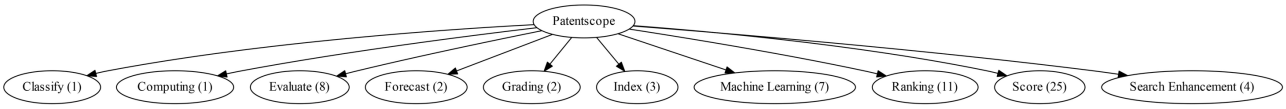

Figure 1. Most frequent innovation type in WIPO's Patentscope.

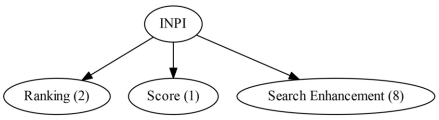

Figure 2. Most frequent innovation type in INPI's patent database.

Score (25 times) is the most frequent type applied in WIPO's Patentscope, which means that a great part of the patent application documents is about scoring methods (approximately 39%). In INPI's patent database, Search Enhancement methods (8 times) were the subject of most patent applications (73% of them), which means that those innovations concern ways to improve search results. Ranking appears as the second most frequent type in patent application documents in both databases, 17% in WIPO's Patentscope and 18% in INPI's patent database.

3.2 What are the commonest International Patent Classifications (IPC) (SRQ2)?

The International Patent Classification (IPC) is a system for classifying and searching patent applications and finding out about specifications of granted patents, utility models, and suchlike technical documents. It is divided into eight sections (A-H) with subdivisions consisting of alphanumeric codes representing hierarchical levels. The great majority of the patent applications studied pertain to the Physics section (represented by letter G), as Figures 3 and 4 show. In both databases, the most frequent used IPC code is G06F 17/30, which stands for G (Physics), 06 (Computing; Calculation or Counting), F (Electrical Digital Data Processing), 17(digital computing or data processing equipment or methods), 30 (information retrieval). As a document can receive more than one IPC code, this study compiled 34 codes at WIPO's Patentscope and ten codes at INPI's patent database.

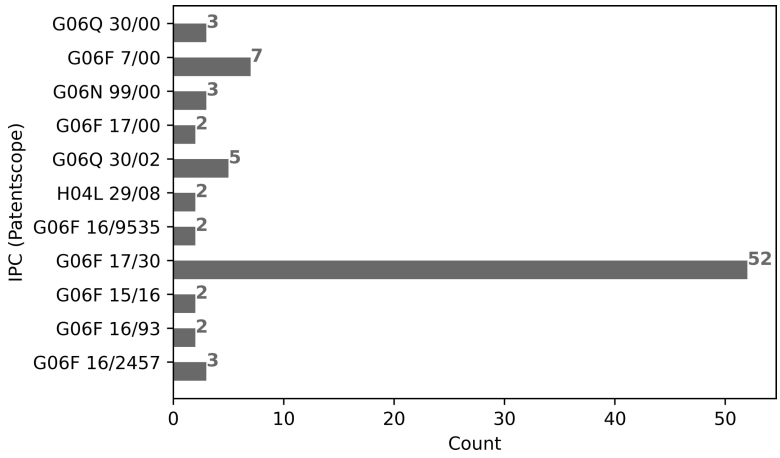

Figure 3. Most frequent IPC code in WIPO's Patentscope.

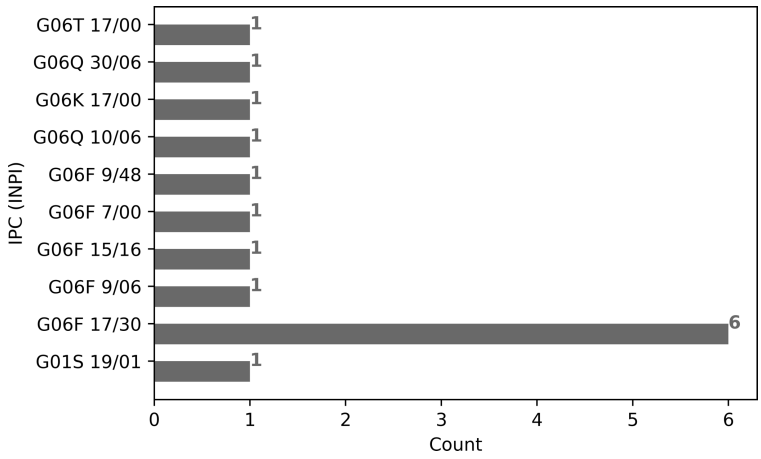

Figure 4. Most frequent IPC code in INPI's patent database.

3.3 Who are the patent holders (SRQ3)?

Since the 1990's an increase in the number and share of academic patents has been observed in the USA and Europe for a variety of reasons, most of them concerning changes in regulation (van Zeebroeck et al. 2008). In 1980 the U.S. Bayh-Dole Act allowed "universities to patent the results of federally-funded research and license the resulting technology to businesses and other entities" (Agrawal and Henderson 2002). Whereas in Europe, more than a few countries adopted analogous legislation (e.g., UK and Germany in 1998, and Belgium in 1999), putting in place some incentive mechanisms such as granting a portion of the royalties derived from their patented discoveries to the researchers themselves (van Zeebroeck et al. 2008). Although private companies seem to be the major beneficiaries of the changes in Europe, as they are the main owners of academic patents, in the United States, the role is taken by universities

(Backs et al. 2019). Nevertheless, not all research fields benefited equally (van Zeebroeck et al. 2008). With this question, we seek to find out if there is a trend of patent applications from academic origins in the search engine industry.

As patent application documents allow more than one applicant (the patent holder) registry, our study encountered 75 applicants in WIPO's Patentscope and 12 in INPI's patent database. Figures 5 and 6 show the results, and Table 8 details the patent documents. Patent applications from academia origins are not common in the search engine industry, about 4% abroad and 8% in Brazil. Therefore, we do not perceive any trend of academic patent applications.

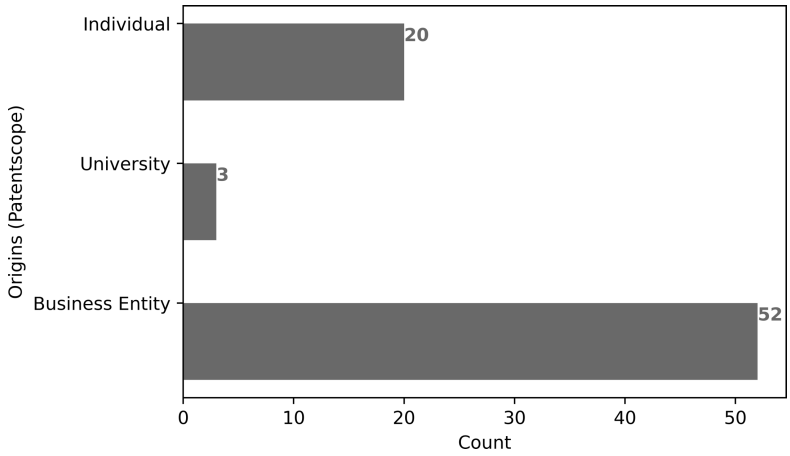

Figure 5. Patent Application Document Origins in WIPO's Patentscope.

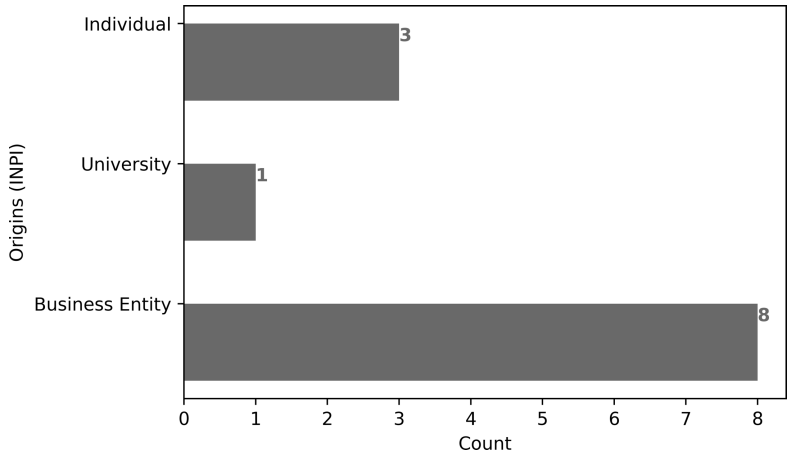

Figure 6. Patent Application Document Origins in INPI's patent database.

We decided to compile the patent holders' names from the applicant information provided in the patent documents to supply a complementary view. As the patent holders sum up hundreds of names, Figures 7 and 8 provide only instances of the most common compiled ones. It is possible to note that private companies file more patents in both databases. In WIPO's Patentscope, Google is responsible for almost 36% of the

applications, with Microsoft as runner-up (12.5%). In INPI's patent database, Microsoft is the top patent holder with 18% of the applications. A total of four academic institutions filed patent documents: Northwestern University, Stanford University, and École Polytechnique Fédérale de Lausanne in WIPO's Patentscope and Università di Firenze in INPI's patent database.

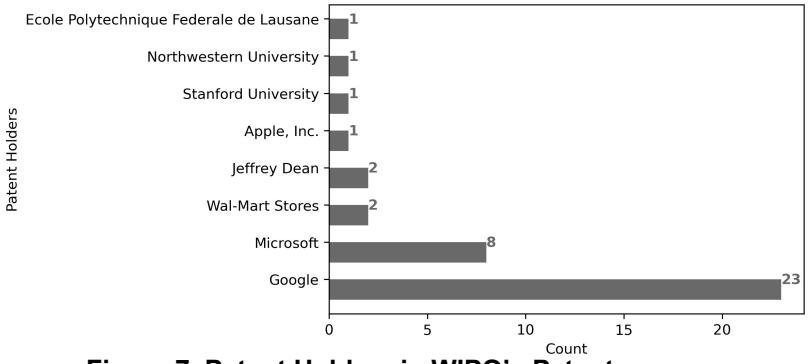

Figure 7. Patent Holders in WIPO's Patentscope.

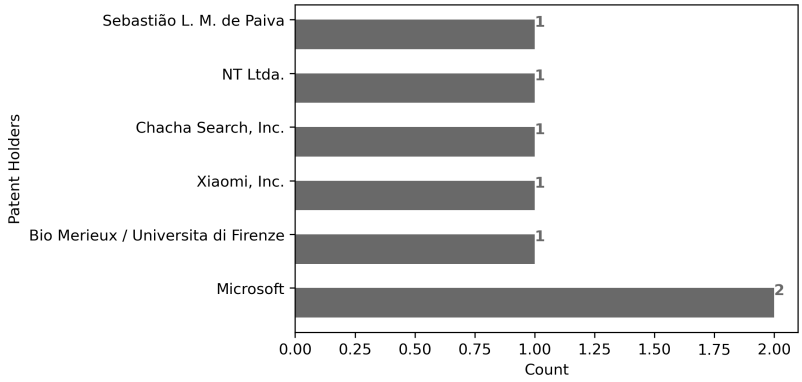

Figure 8. Patent Holders in INPI's patent database.

3.4 From what countries do the patent holders come (SRQ4)?

As the world becomes more interconnected and interdependent, and succeeding in this new economy depends on the prospects of developing and sustaining a knowledge-based society, a new inter-relational system built on openness, flexibility, permanent education and specialized entrepreneurial motivation imposes itself (Lungu 2019). In this context, there is a clear advantage to places in which patents provide incentives to entrepreneurs to engage in further technological development and to introduce useful products or services into the market (Mazzoleni 2006).

For this reason, this particular Research Question intended to provide an overview of the places from which this study's patent application documents originated. The United States appears as the search engine industry's hotspot with 77% of all applications, considering both databases, leading in WIPO's Patentscope and INPI's patent database. Way ahead of the runner-up Brazil, with approximately 5%, which appears solely

in INPI's patent database. The United States also appears in PCT applications and applications to the European Patent Office. Figures 9 and 10 provide the details for each database, and Table 9 shows the patent documents from each country.

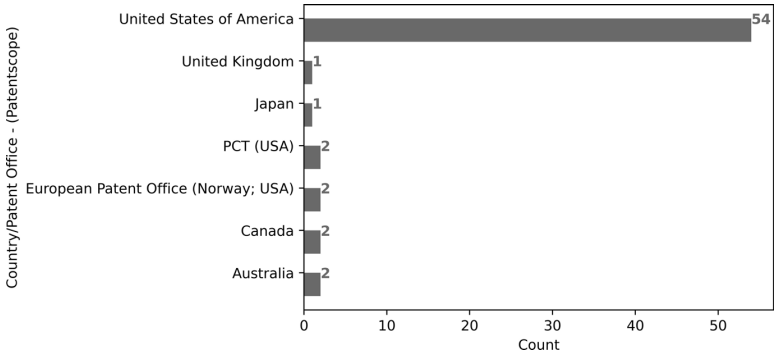

Figure 9. Countries from which the patent application documents came from in WIPO's Patentscope.

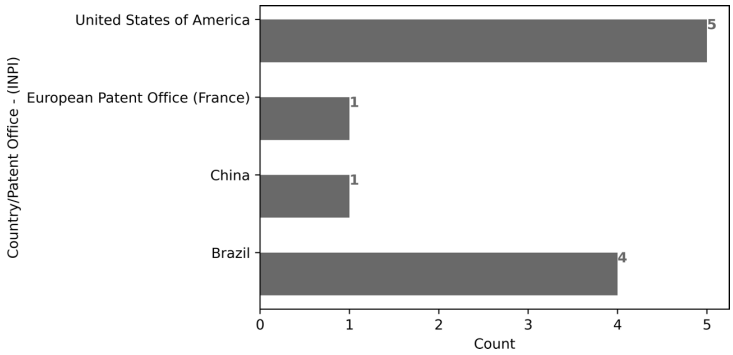

Figure 10. Countries from which the patent application documents came from in INPI's patent database.

3.5 What are the patent documents' application date (SRQ5)?

A technological innovation process is uncertain as it is impossible to know its results beforehand (Procaci et al. 2016). In such a scenario, a patent application can be considered a possible innovation or at least an attempt to it (Procaci et al. 2016). Figures 11 and 12 show a "peaks and valleys" pattern throughout the years, with numbers fluctuating. In mapping the patent applications' evolution, the current Research Question tries to provide a picture regarding the search engine industry's proficiency to churn its innovative landscape. Based on the results, its capability seems unsteady.

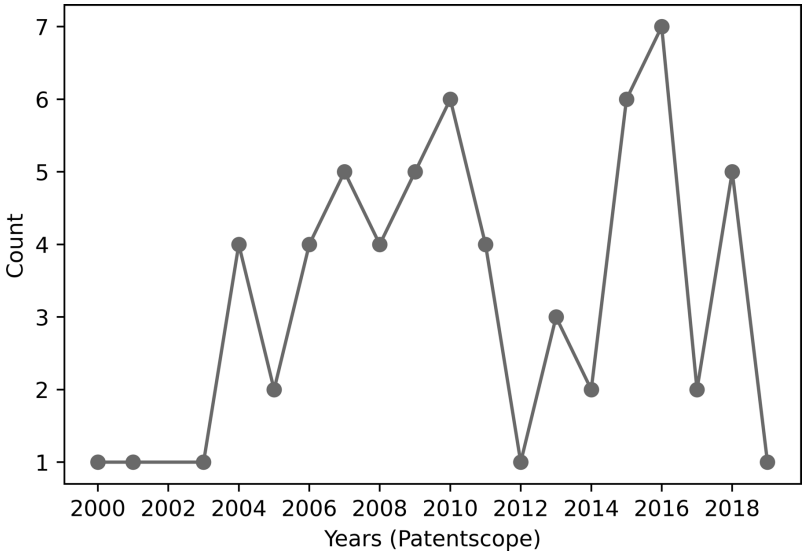

Figure 11. Patent application numbers over time in WIPO's Patentscope.

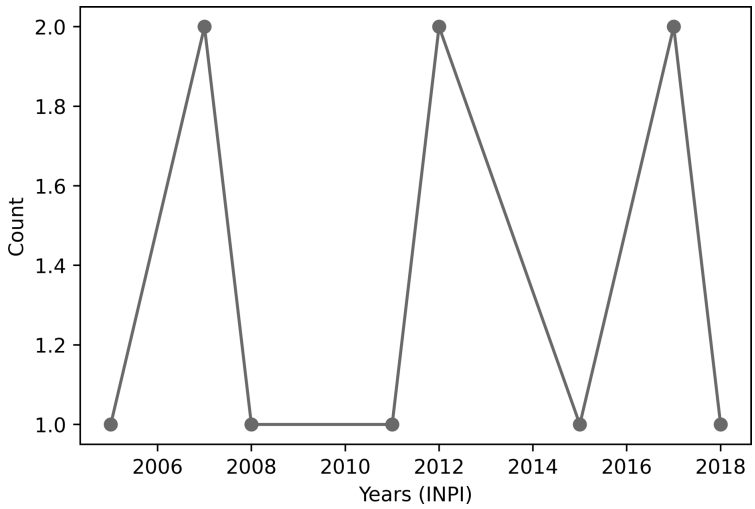

Figure 12. Patent application numbers over time in INPI's patent database.

4 Discussion

Web search engines commonly measure the relevance between a query and Webpages by Text Matching models (Zhang et al. 2019a), also known as pure string matching. In this case, documents are retrieved mainly by matching the terms presented at the query to the documents' content. However, Text Matching models fail to find target information toward the user's intent. Hence, documents containing target information that users wish for may not be ranked at the top positions or even not be summoned up at all (Zhang et al. 2019a). As voice search and voice-powered assistants enter more

and more people's everyday lives, queries tend to be as long tailed[8] (Zhang et al. 2019a). A situation that demands search engines to evolve into a higher level of semantic relevance matching. For this reason, modern search scenarios request intent-target relevance matching models (Zhang et al. 2019a), which brings to the limelight the intent matching paradigm academia now shifts its attention to.

Higher education institutions have an important part in the process of transmitting, producing, and transferring knowledge within a knowledge-based economy (Lungu 2019), with patents as one of its most visible and profitable output (US$28 billion in net revenue in 2015) (Backs et al. 2019). As one of the major state-of-the-art sources, it seems reasonable to believe that academia does have an influential role in setting a technical state-of-the-art agenda. Accordingly, we set our Primary Research Question to find out if academia is influencing commercial innovations regarding the intent matching paradigm and how much of this "innovation footprint" can be mapped. Our PRQ1 summarized this goal as "do search engine industry's innovations reflect the paradigm change detected in academia?".

We needed to identify an intent matching trend in the batch of patent application documents analyzed to answer it. To that end, academic studies regarded as related work under the intent matching paradigm were compiled. We borrow from their research topics to form 8 intent matching indicators used as guidelines to establish its presence in the patent applications. Table 10 presents this related work, showing the intent matching indicators, definition, and references.

Figures 13 and 14 show the results on both databases. All indicators were mapped in WIPO's Patentscope. They were identified in 26 patent application documents out of 64 (in some documents, more than one indicator was found), which accounted for 40% of the hits. In INPI's patent database, two indicators were mapped in 3 out of 11 documents (27%). Table 11 shows all the 29 patent documents identified.

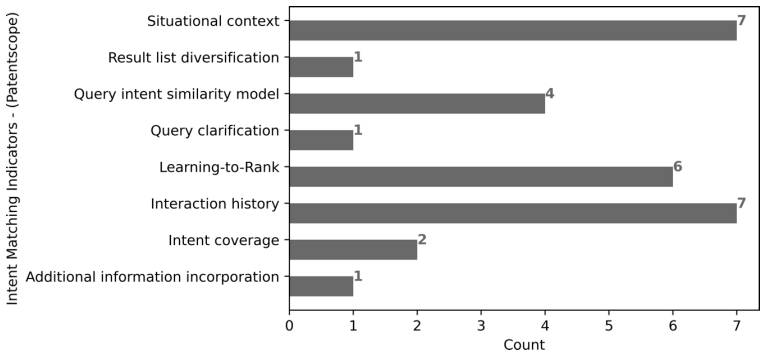

Figure 13. Intent Matching Indicators mapped in WIPO's Patentscope.

[8] In statistics, the term "long tail" refers to a type of distribution in which its major portion has occurrences far from its top (known as the "head") or central part.

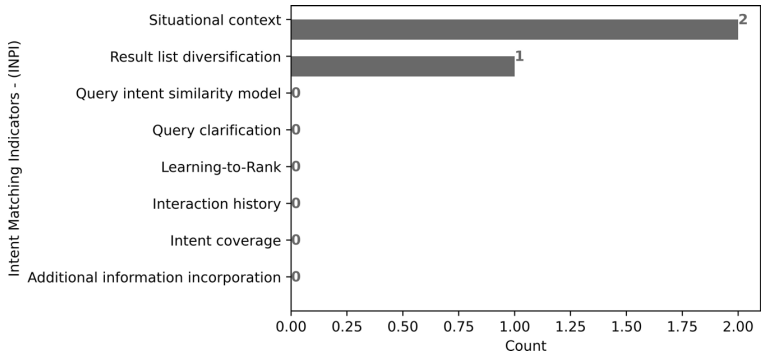

Figure 14. Intent Matching Indicators mapped in INPI's patent database.

Overall, the indicator "situational context" had most hits. The approach relies on measures such as click-through rates[9] (CTR) and click positions[10] across countries, languages, and days of the week to determine search behaviors that can be employed in ranking models (i.e., context-aware ranking model). Grounded on the assumption that if "a user searching behavior may depend on the situational context of a query", then "the properties of the current search request are independent from the query content" (Zamani et al. 2017), meaning that situational features like user device, browser, and language, influence user searching behavior and therefore can be used to fill the gaps in query understanding. Based on the patent application documents mapped, this is the sort of innovation the search engine industry is implementing in 12% of the cases (9 out of 75 patent applications) in Brazil and abroad.

User searching behavior seems to be a focal point to the industry, as "interaction history" demonstrates (9.33% overall) as the second favorite intent matching indicator we encountered in the patent documents. The history of users' interactions is a valuable source of information about their search intent, and it is extensively utilized not only to improve ranking by constructing ranking features but also to personalize search (Ustinovskiy and Serdyukov 2013). It is most used two-fold to extract stable preferences of a user and to specify or to disambiguate the current query.

Artificial Intelligence cuts the "top 3" employing the machine learning technique Learning-to-Rank (8% overall). The search engine industry seems to be attending the ranking problem[11] of predicting the relevance labels *yi* of a query/document pair (*q/di*) by computing behavior patterns related to query intention. The "query intent similarity model" seems to corroborate the reasoning (5.33% overall), as it appears in the analyzed patent application documents as the fourth most frequent intent matching indicator. The approach groups query with similar intent together (e.g., "cheap cars" and "low-priced autos") and attributes labels to their intent similarities (Zhang et al. 2019b), which could be used as relevance labels to Learning-to-Rank.

[9] The percentage of people accessing a hypertext link.

[10] Concerns the numeric position value of a search result element describing its position on the page.

[11] It is a classical problem in Information Retrieval. It concerns ranking the query results according to a relevance-based criterion.

Based on the discussed results, the authors are confident to assert that a significant part of the search engine industry's innovations analyzed in this systematic mapping study (40% in WIPO's Patentscope and 27% in INPI's patent database) do reflect the paradigm change from pure string matching to intent matching detected in academia.

5 Threats to validity

This systematic mapping study is subjected to three types of threats: to its construct validity (research design), to its internal validity (data extraction), and to its conclusion validity (reliability). We consider that threats to its external validity (generalizability) are irrelevant as the results cannot be generalized to other problem domains (i.e., Information Retrieval) or other solution domains (i.e., intent matching).

About threats concerning construct validity (research design), the results presented in Sections 3 and 4 are valid only to our sample of patent application documents. Therefore, we tried to ensure the inclusion of as many relevant documents as possible with the designed inclusion criteria and the quality questions checklist. During the initial screening, only the title and abstract were considered. To prevent excluding relevant patent application documents based on the lack of investigation, we decided to include documents we were uncertain of temporarily. In the subsequent classification, the claim and specification sections were read, and inclusion or exclusion was decided. Still, as it is ultimately a matter of subjective choice, there is always a possibility of a poor choice.

Threats to internal validity might have arisen from the data extraction. As the study included only two databases, the retrieved documents were restricted to WIPO's Patentscope and INPI's patent database. In this case, relevant patent applications from other databases might not be included. Also, the fact that the databases' searching characteristics prevented us from applying the very same search string can be considered a limitation.

Threats to conclusion validity might have arisen from wrong conclusions and the study's replicability. Concerning the former, we have discussed issues that could lead to wrong conclusions in the contexts of threats to construct validity and internal validity. They still hold in this case. Regarding the latter, we tried to mitigate it by detailing the complete method in Section 2. Though we consider that it enables replicating every phase of this mapping study, the patent application documents are still limited to the timeframe in which this study was conducted, namely, September and October 2020 and January 2022.

6 Final Remarks

In this report, we presented a systematic mapping study intended to characterize the technical state-of-the-art in patent documents of methods employed to score documents and search queries, and to grade the search in search engines. We also intended to verify if academia had some role in setting a technical state-of-the-art agenda for commercial innovation ventures in the search engine industry. The study showed that academia does influence the search engine industry's innovations, at least concerning

the paradigm change from pure string matching to intent matching. 40% of the patent application documents considered suitable in WIPO's Patentscope and 27% in INPI's patent database had indicators regarding intent matching topics. Concerning the former goal, we found out that a great part of the patent application documents is about scoring methods (39% in WIPO's Patentscope) or methods to improve search results (73% in INPI's patent database). Also, 4% of the patent application documents in WIPO's Patentscope and 8% in INPI's patent database have academic origins, which indicates that although academia does influence the search engine industry's technical state-of-the-art, there is no trend of patent applications filed by academic institutions. Besides, most patent documents (77% overall) come from the United States. Finally, we discovered that the patent application numbers in the search engine industry fluctuate throughout the years, showing a "peaks and valleys" pattern. It is a sign that this industry's proficiency to churn its innovative landscape is, at best, unsteady.

Based on the patent application trends observed, we conclude this report by discussing some research directions that would possibly involve a multidisciplinary approach. Firstly, we think it is worth further investigating why few patent applications are originated from academia, especially in Brazil. Other studies had hinted to pressures for academic publishing by the institutions that regulate science in the country or to slowness and bureaucracy in establishing partnerships between business companies and universities (Procaci et al. 2016). Although it might be the case, the gap between the application and grant date observed in the studied patent documents (sometimes years-long span) seems discouraging to hold a scientific communication. Secondly, one of our quality questions (QQ2) made us investigate the patent application documents for insights regarding what would infringe the patent. With vanishing frontiers between basic and applied research (van Zeebroeck et al. 2008), it seems possible that academic institutions would find themselves more likely to infringe patents. In this case, it would be interesting to study how this important issue regarding a knowledge-based society impacts research in Brazil and if research exemptions policies would be necessary. Thirdly, it would be interesting to find out more about the virtuous relationship between academia and the search engine industry agendas. For example, do granted patents (which were outside the scope of this study) enhance incentives for further research and development efforts in both academia and business industry?

Table 1. Research Questions.

Primary Research Question	
Research Question	**Data to be extracted**
PRQ1: Do the search engine industry's innovations reflect the paradigm change detected in academia outputs?	The patent application document's claims and specification sections, whose descriptions will be used as an "innovation footprint" compared to research topics from 8 academic papers regarded as related work under the intent matching paradigm. The academic papers are (Santos et al. 2015); (Ustinovskiy and Serdyukov 2013); (Zamani et al. 2017); (Zhang et al. 2019b); (Kiesel et al. 2018); (Yu et al. 2018); (Roy et al. 2020); (Lucchese et al. 2015).

Secondary Research Questions	
Research Question	**Data to be extracted**
SRQ1: What are the inventions that the search engine industry most frequently claim as technological developments in the patent documents?	Compilation of the innovation claims from the patent documents.
SRQ2: What are the commonest International Patent Classifications (IPC)[a]?	The International Patent Classification (IPC).
SRQ3: Who are the patent holders?	Individual, business entity or University that filed the patent document.
SRQ4: From what countries do the patent holders come?	Compilation of the countries from which patent holders came. If an application was filed via PCT[b]or EPC[c], then the country from which it was applied will be identified.
SRQ5: What are the patent documents' application date?	The year in which the patent document was filed.

[a] The International Patent Classification (IPC) is used to classify the patent document's technical content.

[b] Patent Cooperation Treaty (PCT) is the closest instance to an international patent system, and it is composed of around 137 countries. Patent applications can be divided into two phases: international and national. The international is composed of an international application, a mandatory search and a preliminary examination (the latter is optional). The research results and an opinion are then issued and will serve as a reference to both the patent feasibility for those applying it and the national offices, which are part of the second phase. Although the system facilitates patent applications in each member nation, it is important to note that it does not eliminate the need to formulate the application in each national office.

[c] European Patent Convention (EPC) is a multilateral treaty created to provide a unified patent system to Europe. EPC is a partnership between the European Patent Office (EPO) and the national offices of the participating nations in the European Union. The patent would be secured across Europe through the European Patent Office.

Table 2. Scope of the study defined using PICOC framework.

PICOC	Term
Population	search engine
Intervention	score document, grade the search
Comparison	-
Outcome	method, algorithm, computing, systems, apparatus
Context	information retrieval

Table 3. Search string defined to identify relevant documents.

English string
(″search engine″) AND (″grade the search″OR ″score document″) AND (″algorithm″OR ″apparatus″OR ″computing″OR ″method″OR ″systems″)
Portuguese version
(″ferramenta de busca″) AND (″avaliac¸a˜o da busca″OR ″pontuac¸a˜o do documento″) AND (″algoritmo″OR ″dispositivo″OR ″computac¸a˜o″OR ″me´todo″OR ″sistemas″)

Table 4. The search strings applied to INPI's patent database.

Search Strings in Portuguese	**English Version**
(″ferramenta de busca″) AND (″algoritmo″)	(″search engine″) AND (″algorithm″)
(″ferramenta de busca″) AND (″dispositivo″)	(″search engine″) AND (″apparatus″)
(″ferramenta de busca″) AND (″computac¸a˜o″)	(″search engine″) AND (″computing″)
(″ferramenta de busca″) AND (″me´todo″)	(″search engine″) AND (″method″)
(″ferramenta de busca″) AND (″sistema″)	(″search engine″) AND (″systems″)

Table 5. Quality Assessment Checklist.

Quality Questions (QQ)	**Scores**
QQ1: Does the abstract provide a summary of the invention to aid searching?	Yes (1pt) Partially (0.5pt) No (0pt)
QQ2: Does the claim section set out the essential features of the invention in a manner to clearly define what would infringe the patent?	Yes (1pt) Partially (0.5pt) No (0pt)
QQ3: Was the document describing the invention for which the patent is sought appropriate?	Yes (1pt) Partially (0.5pt) No (0pt)
QQ4: Was the innovation suitable for the research objective?	Yes (1pt) Partially (0.5pt) No (0pt)
QQ5: Was the patent application under the Patent Cooperation Treaty (PCT)?	Yes (1pt) No (0pt)

Table 6. Patent documents' selection details.

Database	Patent documents retrieved from the search string	Patent documents identified as repeated	Patent documents evaluated by title and abstract reading	Patent documents considered suitable after Inclusion and Exclusion Criteria	Patent documents considered suitable after the Quality Questions
Patentscope	675	105	570	64	64
INPI	38	16	22	11	11

Table 7. Patent Documents Selected.

Patent Documents from WIPO's Patentscope
WPS1 (US20080215574), WPS2 (US20160098488), WPS3 (US20190294692), WPS4 (EP1930816), WPS5 (US20130282707), WPS6 (US20110022600), WPS7 (US08799107), WPS8 (US20060059138), WPS9 (US20180217991), WPS10 (US20160321363), WPS11 (CA2385570), WPS12 (US20070185847), WPS13 (US08145618), WPS14 (US20170177720), WPS15 (US20060149723), WPS16 (US20120130994), WPS17 (US08195651), WPS18 (US7587391), WPS19 (US20180107940), WPS20 (CA2538597), WPS21 (US09563692), WPS22 (EP2416262), WPS23 (US20110264647), WPS24 (US20120095994), WPS25 (US20110264518), WPS26 (US20160283474), WPS27 (US7523099), WPS28 (US20120089588), WPS29 (US20120011117), WPS30 (US08244722), WPS31 (AU2007333558), WPS32 (US09002867), WPS33 (JP2007128547), WPS34 (US09836461), WPS35 (US08504411), WPS36 (US09483568), WPS37 (US20170024390), WPS38 (US20170039267), WPS39 (US20180300315), WPS40 (US20180046716), WPS41 (GB2472250), WPS42 (AU2007200526), WPS43 (US20120209838), WPS44 (US20180300410), WPS45 (US10394830), WPS46 (US20070088693), WPS47 (US08463774), WPS48 (US20190205472), WPS49 (US09286387), WPS50 (US08849830), WPS51 (US20170091319), WPS52 (US20200210438), WPS53 (US20180165288), WPS54 (US20090228777), WPS55 (US20200104946), WPS56 (US20020198869), WPS57 (US7283997), WPS58 (US20160283489), WPS59 (US20160170984), WPS60 (WO2010021723), WPS61 (US08255386), WPS62 (WO2020081082), WPS63 (US20080250105), WPS64 (US20160012056).
Patent Documents from INPI's patent database
IPD1 (BR 11 2014 016329 4 A8), IPD2 (BR 11 2013 031948 8 A2), IPD3 (BR 11 2015 019848 1 A2), IPD4 (PI 0802762-5 A2), IPD5 (PI 0707294-5 A2), IPD6 (BR 11 2012 028553 0 A2), IPD7 (PI 0706683-0 A2), IPD8 (PI 0506009-5 A2), IPD9 (BR 10 2017 000526 7 A2), IPD10 (BR 10 2017 018178 2 A2), IPD11 (BR 10 2018 000113 2 A2).

Table 8. Patent Documents from both databases detailed per origin.

Individual	WPS7 (US08799107), WPS14 (US20170177720), WPS16 (US20120130994), WPS19 (US20180107940), WPS24 (US20120095994), WPS25 (US20110264518), WPS29 (US20120011117), WPS32 (US09002867), WPS35 (US08504411), WPS41 (GB2472250), WPS43 (US20120209838), WPS44 (US20180300410), WPS47 (US08463774), WPS49 (US09286387), WPS50 (US08849830), WPS58(US20160283489), WPS59 (US20160170984), WPS60 (WO2010021723), WPS63 (US20080250105), WPS64 (US20160012056), IPD4 (PI 0802762-5 A2), IPD9 (BR 10 2017 000526 7 A2), IPD10 (BR 10 2017 018178 2 A2)
University	WPS6 (US20110022600), WPS12 (US20070185847), WPS17 (US08195651), IPD1 (BR 11 2014 016329 4 A8)
Business Entity	WPS1 (US20080215574), WPS2 (US20160098488), WPS3 (US20190294692), WPS4 (EP1930816), WPS5 (US20130282707), WPS7 (US08799107), WPS8 (US20060059138), WPS9 (US20180217991), WPS10 (US20160321363), WPS11 (CA2385570), WPS12 (US20070185847), WPS13 (US08145618), WPS15 (US20060149723), WPS16 (US20120130994), WPS18 (US7587391), WPS20 (CA2538597), WPS21 (US09563692), WPS22 (EP2416262), WPS23 (US20110264647), WPS25 (US20110264518), WPS26 (US20160283474), WPS27 (US7523099), WPS28 (US20120089588), WPS29 (US20120011117), WPS30 (US08244722), WPS31 (AU2007333558), WPS32 (US09002867), WPS33 (JP2007128547), WPS34 (US09836461), WPS35 (US08504411), WPS36 (US09483568), WPS37 (US20170024390), WPS38 (US20170039267), WPS39 (US20180300315), WPS40 (US20180046716), WPS42 (AU2007200526), WPS43 (US20120209838), WPS45 (US10394830), WPS46 (US20070088693), WPS47 (US08463774), WPS48 (US20190205472), WPS49 (US09286387), WPS50 (US08849830), WPS51 (US20170091319), WPS52 (US20200210438), WPS53 (US20180165288), WPS54 (US20090228777), WPS55 (US20200104946), WPS56 (US20020198869), WPS57 (US7283997), WPS61 (US08255386), WPS62 (WO2020081082), IPD1 (BR 11 2014 016329 4 A8), IPD2 (BR 11 2013 031948 8 A2), IPD3 (BR 11 2015 019848 1 A2), IPD5 (PI 0707294-5 A2), IPD6 (BR 11 2012 028553 0 A2), IPD7 (PI 0706683-0 A2), IPD8 (PI 0506009-5 A2), IPD11 (BR 10 2018 000113 2 A2)

Table 9. Patent Documents per country.

Country	Patent Document
USA	WPS1 (US20080215574), WPS2 (US20160098488), WPS3 (US20190294692), WPS5 (US20130282707), WPS6 (US20110022600), WPS7 (US08799107), WPS8 (US20060059138), WPS9 (US20180217991), WPS10 (US20160321363), WPS12 (US20070185847), WPS13 (US08145618), WPS14 (US20170177720), WPS15 (US20060149723), WPS16 (US20120130994), WPS17 (US08195651), WPS18 (US7587391), WPS19 (US20180107940), WPS21 (US09563692), WPS22 (EP2416262), WPS23 (US20110264647), WPS24 (US20120095994), WPS25 (US20110264518), WPS26 (US20160283474), WPS27 (US7523099), WPS28 (US20120089588), WPS29 (US20120011117), WPS30 (US08244722), WPS32 (US09002867), WPS34 (US09836461), WPS35 (US08504411), WPS36 (US09483568), WPS37 (US20170024390), WPS38 (US20170039267), WPS39 (US20180300315), WPS40 (US20180046716), WPS43 (US20120209838), WPS44 (US20180300410), WPS45 (US10394830), WPS46 (US20070088693), WPS47 (US08463774), WPS48 (US20190205472), WPS49 (US09286387), WPS50 (US08849830), WPS51 (US20170091319), WPS52 (US20200210438), WPS53 (US20180165288), WPS54 (US20090228777), WPS55 (US20200104946), WPS56 (US20020198869), WPS57 (US7283997), WPS58 (US20160283489), WPS59 (US20160170984), WPS60 (WO2010021723), WPS61 (US08255386), WPS62 (WO2020081082), WPS63 (US20080250105), WPS64 (US20160012056), IPD2 (BR 11 2013 031948 8 A2), IPD5 (PI 0707294-5 A2), IPD6 (BR 11 2012 028553 0 A2), IPD7 (PI 0706683-0 A2), IPD8 (PI 0506009-5 A2)
Brazil	IPD4 (PI 0802762-5 A2), IPD9 (BR 10 2017 000526 7 A2), IPD10 (BR 10 2017 018178 2 A2), IPD11 (BR 10 2018 000113 2 A2)
Canada	WPS11 (CA2385570), WPS20 (CA2538597)
Australia	WPS31 (AU2007333558), WPS42 (AU2007200526)
China	IPD3 (BR 11 2015 019848 1 A2)
Norway	WPS4 (EP1930816)
France	IPD1 (BR 11 2014 016329 4 A8)
UK	WPS41 (GB2472250)
Japan	WPS33 (JP2007128547)

Table 10. Intent Matching Indicators.

Intent Matching Indicators	Definition	Reference
Result list diversification	Diversify the search result list by updating the relevance ranking through the identification of documents that carried different information from those documents already seen by the user.	(Santos et al. 2015)
Interaction history	The extraction of information from users' search and browsing long-term history and the usage of short-term history to determine the context of a given query.	(Ustinovskiy and Serdyukov 2013)
Situational context	Improvement of ranking performance by using the contextual features of the current search request that are independent from both query content and user history. For example, situational context can depend on search request time and location.	(Zamani et al. 2017)
Intent coverage	Encoding model to identify previously observed queries with the same search intent.	(Zhang et al. 2019b)
Query clarification	Automatic aiding approach which helps users to formulate their search intent by suggestions that help subsequent query modeling and query understanding tasks.	(Kiesel et al. 2018)
Query intent similarity model	Approach to evaluate the quality of intent representations by their ability to group queries with similar in tents together.	(Zhang et al. 2019b)
Additional information incorporation	Approach to evaluate the relevance of an attribute by calculating the amount of information gained about a variable from observing another variable.	(Yu et al. 2018) and (Roy et al. 2020)
Learning-to-Rank	Ranking query results according to a relevance criterion using machine learning techniques to score a set of candidate documents according to their relevance to a given user's query.	(Lucchese et al. 2015)

Table 11. Patent Documents identified by the Intent Matching Indicators.

Intent Matching Indicators	Patent Documents
Result list diversification	WPS28 (US20120089588), IPD7 (PI 0706683-0 A2)
Interaction history	WPS2 (US20160098488), WPS8 (US20060059138), WPS18 (US7587391), WPS22 (EP2416262), WPS33 (JP2007128547), WPS42 (AU2007200526), WPS63 (US20080250105)
Situational context	WPS7 (US08799107), WPS12 (US20070185847), WPS15 (US20060149723), WPS20 (CA2538597), WPS24 (US20120095994), WPS26 (US20160283474), WPS31 (AU2007333558), IPD10 (BR 10 2017 018178 2 A2), IPD11 (BR 10 2018 000113 2 A2)
Intent coverage	WPS19 (US20180107940), WPS29 (US20120011117)
Query clarification	WPS21 (US09563692)
Query intent similarity model	WPS3 (US20190294692), WPS19 (US20180107940), WPS23 (US20110264647), WPS57 (US7283997)
Additional information incorporation	WPS62 (WO2020081082)
Learning-to-Rank	WPS3 (US20190294692), WPS9 (US20180217991), WPS24 (US20120095994), WPS25 (US20110264518), WPS39 (US20180300315), WPS53 (US20180165288)

References

[Agrawal and Henderson 2002] Agrawal, A. and Henderson, R. (2002). Putting patents in context: Exploring knowledge transfer from MIT. *Management Science*, 48(1):44–60.

[Backs et al. 2019] Backs, S., Gu¨nther, M., and Stummer, C. (2019). Stimulating academic patenting in a university ecosystem: an agent-based simulation approach. *The Journal of Technology Transfer*, 44(2):434–461.

[Buganza and Della Valle 2010] Buganza, T. and Della Valle, E. (2010). Chapter 4: The search engine industry. In Ceri, S. and Brambilla, M., editors, *Search Computing: Challenges and Directions*, pages 45–71. Springer Berlin Heidelberg.

[Clarke 2009] Clarke, N. S. (2009). Searching patents. In Jolly, A. and Philpott, J., editors, *The Handbook of European Intellectual Property Management: Developing, Managing and Protecting Your Company's Intellectual Property*, chapter 3.1, pages 117–121. Kogan Page.

[Clarke 2018] Clarke, N. S. (2018). The basics of patent searching. *World Patent Information*, 54: S4–S10.

[Dziallas and Blind 2019] Dziallas, M. and Blind, K. (2019). Innovation indicators throughout the innovation process: An extensive literature analysis. *Technovation*, 80-81: 3–29.

[Khattak et al. 2020] Khattak, A., Habib, A., Asghar, M. Z., Subhan, F., Razzak, I., and Habib, A. (2020). Applying deep neural networks for user intention identification. *soft computing*, pages 1–30.

[Kiesel et al. 2018] Kiesel, J., Bahrami, A., Stein, B., Anand, A., and Hagen, M. (2018). Toward voice query clarification. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, SIGIR '18, page 1257–1260, New York, NY, USA. Association for Computing Machinery.

[Lucchese et al. 2015] Lucchese, C., Nardini, F. M., Orlando, S., Perego, R., Tonellotto, N., and Venturini, R. (2015). Quickscorer: A fast algorithm to rank documents with additive ensembles of regression trees. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 73– 82.

[Lungu 2019] Lungu, V. (2019). Knowledge-based society - a condition to ensure sustainable development. Eastern European Journal for Regional Studies (EEJRS), 5(1):96–111.

[Mazzoleni 2006] Mazzoleni, R. (2006). The Effects of University Patenting and Licensing on Downstream R&D Investment and Social Welfare. *The Journal of Technology Transfer*, 31(4):431–441.

[Mohammadi et al. 2020] Mohammadi, S., Chapon, M., and Frémond, A. (2020). Query intent detection from the SEO perspective. In *European Conference on Advances in Databases and Information Systems*, pages 49–59.

[Petersen et al. 2008] Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pages 1–10.

[Procaci et al. 2016] Procaci, T., Araujo, R., Siqueira, S., and Nunes, B. (2016). Prospecc¸a˜o tecnolo´gica: Levantamento de patentes, atuac¸a˜o da academia e potenciais inovac¸o˜es em ambientes de aprendizagem no brasil de 2000 a 2015. *iSys - Revista Brasileira de Sistemas de Informac¸a˜o*, 9 (3):69–88.

[Roy et al. 2020] Roy, N., Moraes, F., and Hauff, C. (2020). Exploring users' learning gains within search sessions. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, page 432–436, New York, NY, USA. Association for Computing Machinery.

[Santos et al. 2015] Santos, R. L. T., MacDonald, C., and Ounis, I. (2015). Search result diversification. *Found. Trends Inf. Retr.*, 9:1–90.

[Silva et al. 2018a] Silva, I., Nunes, M., Felizardo, K. R., Nakagawa, E., Ferrari, F., Fabbri, S., and Santos Ju´nior, J. (2018a). *Almanaque para popularizac¸a˜o de cieˆncia da computac¸a˜o*. SBC, Se´rie 6, Metodologia cient´ıfica e tecnolo´gica; v 7.

[Silva et al. 2018b] Silva, I., Nunes, M., Felizardo, K. R., Nakagawa, E., Ferrari, F., Fabbri, S., and Santos Ju´nior, J. (2018b). *Almanaque para popularizac¸a˜o de cieˆncia da computac¸a˜o*. SBC, Se´rie 6, Metodologia cient´ıfica e tecnolo´gica; v 8.

[Ustinovskiy and Serdyukov 2013] Ustinovskiy, Y. and Serdyukov, P. (2013). Personalization of web-search using short-term browsing context. In *Proceedings of the 22nd ACM International Conference on Information amp; Knowledge Management*, CIKM '13, page 1979–1988, New York, NY, USA. Association for Computing Machinery.

[van Zeebroeck et al. 2008] van Zeebroeck, N., van Pottelsberghe de la Potterie, B., and Guellec, D. (2008). Patents and academic research: a state of the art. *Journal of Intellectual Capital*, 9(2):246–263.

[Wohlin et al. 2012] Wohlin, C., Runeson, P., Ho¨st, M., Ohlsson, M. C., Regnell, B., and Wessle´n, A. (2012). *Experimentation in Software Engineering*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[World Intellectual Property Organization 2019] World Intellectual Property Organization (2019). *Inside WIPO*. World Intellectual Property Organization (WIPO), Geneva, Switzerland. `https://www.wipo.int/about-wipo/en/`, Last accessed on 2020-11-16.

[Wortmann et al. 2017] Wortmann, A., Combemale, B., and Barais, O. (2017). A systematic mapping study on modeling for industry 4.0. In *2017 ACM/IEEE*

20th International Conference on Model Driven Engineering Languages and Systems (MODELS), pages 281–291.

[Yu et al. 2018] Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., and Dietze, S. (2018). Predicting user knowledge gain in informational search sessions. In The 41st International ACM SIGIR Conference on Research Development in Information Retrieval, SIGIR '18, page 75–84, New York, NY, USA. Association for Computing Machinery.

[Zamani et al. 2017] Zamani, H., Bendersky, M., Wang, X., and Zhang, M. (2017). Situational context for ranking in personal search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1531–1540, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Zhang et al. 2020] Zhang, B., Qi, X., and Han, X. (2020). An advanced user intent model based on user learning process. *International Journal of Pattern Recognition and Artificial Intelligence*, 34(9):2050024.

[Zhang et al. 2019a] Zhang, C., Wang, H., Zhou, L., Wang, Y., and Chen, C. (2019a). Machine comprehension-incorporated relevance matching. In Wang, J., Shim, K., and Wu, X., editors, *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 856–865. IEEE.

[Zhang et al. 2019b] Zhang, H., Song, X., Xiong, C., Rosset, C., Bennett, P. N., Craswell, N., and Tiwary, S. (2019b). Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 65–74, New York, NY, USA. Association for Computing Machinery.

Appendix 1 Patent Documents' Reference List

Table 12. Patent Document List from WIPO's Patentscope.

Code	WIPO's Patentscope
WPS1	Lin, Chenxi et al. Efficient Retrieval Algorithm by Query Term Discrimination. Applicants: Microsoft Corporation. Agents: Lee & Hayes, PLLC. US20080215574. Application Date: 27.02.2008. Grant Date: 12.04.2011.
WPS2	Battle, Alexis J. et al. Deriving and Using Interaction Profiles. Applicants: Google LLC. Agents: Fish & Richardson P.C. US20160098488. Application Date: 11.12.2015. Grant Date: 20.08.2019.
WPS3	Zhao, Rongkai et al. Ranking and Presenting Search Engine Results Based on Category-specific Ranking Models. Applicants: Home Depot Product Authority, LLC. US20190294692. Application Date: 23.03.2018.
WPS4	Olsen, Oeystein Haug. Contextual Relevance-weighted Result Set Navigation for Search Engines. Applicants: Fast Search & Transfer Asa. EP1930816. Application Date: 06.11.2007.
WPS5	Stiffelman, Oscar B et al. Two-step Combiner for Search Result Scores. Applicants: Discovery Engine Corporation. US20130282707. Application Date: 23.04.2013.
WPS6	Sathe, Saket et al. Method of Data Retrieval, And Search Engine Using Such a Method. Applicants: École Polytechnique Fédérale de Lausanne EPFL. US20110022600. Application Date: 22.07.2009.
WPS7	Dean Jeffrey A. et al. Systems and Methods for Scoring Documents. Applicants: Google Inc. et al. Agents: Harrity & Harrity, LLP. US08799107. Application Date: 30.09.2004. Grant Date: 05.08.2014.
WPS8	Milic-Frayling, Natasa et al. Facility For Highlighting Documents Accessed Through Search or Browsing. Applicants: Microsoft Corporation. Agents: Hope Baldauff Hartman, LLC. US20060059138. Application Date: 21.10.2005. Grant Date: 09.02.2010.
WPS9	Dato, Domenico et al. A Method to Rank Documents by a Computer, Using Additive Ensembles of Regression Trees and Cache Optimisation, and Search Engine Using Such a Method. Applicants: Istella S.p.A. US20180217991. Application Date: 17.06.2015.
WPS10	Erera, Shai et al. Stability Score Based Re-ranking of Search Results. Applicants: International Business Machines Corporation. Agents: Daniel Kligler. US20160321363. Application Date: 29.04.2015. Grant Date: 23.10.2018.
WPS11	Wheeler, David B. et al. System and Method for Performing Similarity Searching. Applicants: Infoglide Corporation. CA2385570. Application Date: 19.09.2000.

Code	WIPO's Patentscope
WPS12	Budzik, Jerome Louis et al. Filtering Context-Sensitive Search Results. Applicants: Perfect Market, Inc.; Northwestern University. Agents: K&L Gates, LLP. US20070185847. Application Date: 31.01.2007.
WPS13	Pfleger, Karl et al. System and Method for Determining a Composite Score For Categorized Search Results. Applicants: Google Inc. Agents: Fish & Richardson P.C. US08145618. Application Date: 11.10.2010. Grant Date: 27.03.2012.

Table 13. Patent Document List from WIPO's Patentscope.

Code	WIPO's Patentscope
WPS14	Peters, Richard R.; Karmarkar, Amit. Method And System of Scoring Documents Based on Attributes Obtained from a Digital Document By Eye-tracking Data Analysis. Applicants: Richard R Peters; Amit Karmarkar. US20170177720. Application Date: 27.01.2014. Grant Date: 08.08.2017.
WPS15	Finger, II James Charles. System And Method for Providing Search Results with Configurable Scoring Formula. Applicants: Microsoft Corporation. Agents: Woodcock Washburn LLP. US20060149723. Application Date: 06.03.2006. Grant Date: 17.03.2009.
WPS16	Risvik, Knut Magne et al. Matching Funnel for Large Document Index. Applicants: Microsoft Corporation et al. Agents: Shook Hardy & Bacon LLP. US20120130994. Application Date: 22.11.2010. Grant Date: 31.12.2013.
WPS17	Page, Lawrence. Scoring Documents in a Linked Database. Applicants: The Board of Trustees of the Leland Stanford Junior University. Agents: Harrity & Harrity, LLP. US08195651. Application Date: 02.02.2010. Grant Date: 05.06.2012.
WPS18	Bostock, Michael C.; Wu, Alexander C. Method and Apparatus for Generating a Preference Ranking. Applicants: Google Inc. Agents: Park, Vaughan & Fleming LLP. US7587391. Application Date: 13.06.2006. Grant Date: 08.09.2009
WPS19	Lieberman, Jeremy. Artificial Intelligence Method and Apparatus. Applicants: Jeremy Lieberman. US20180107940. Application Date: 03.03.2014.
WPS20	Tong, Simon; Pearson, Mark. Methods And Systems for Improving a Search Ranking Using Population Information. Applicants: Google Inc. Agents: Fish & Richardson P.C. CA2538597. Application Date: 10.09.2004. Grant Date: 31.07.2012.
WPS21	Haahr, Paul; Martin, Charles E. Providing Result-based Query Suggestions. Applicants: Google Inc. Agents: Fish & Richardson P.C. US09563692. Ap plication Date: 24.04.2015. Grant Date: 07.02.2017.
WPS22	Acharya, Anurag et al. Information Retrieval Based on Historical Data. Applicants: Google Inc. EP2416262. Application Date: 15.09.2004.

Code	WIPO's Patentscope
WPS23	Lu, Yumao et al. Query Processing for Web Search. Applicants: Yahoo! Inc. Agents: Berkeley Law & Technology Group, LLP. US20110264647. Application Date: 01.07.2011. Grant Date: 23.10.2012.
WPS24	Nagendra, Nagarajayya. Intelligent Search Appliance with Memory and Feedback. Applicants: Nagarajayya Nagendra. US20120095994. Application Date: 16.10.2011.
WPS25	Liu, Chao; Wang, Yi-Min. Learning a Ranker to Rank Entities with Automatically Derived Domain-specific Preferences. Applicants: Microsoft Corporation et al. Agents: Medley, Behrens & Lewis, LLC. US20110264518. Application Date: 22.04.2010. Grant Date: 19.03.2019.

Table 14. Patent Document List from WIPO's Patentscope.

Code	WIPO's Patentscope
WPS26	Patterson, Anna L. Multiple Index Based Information Retrieval System. Applicants: Google Inc.; Google LLC. Agents: Brake Hughes Bellermann LLP. US20160283474. Application Date: 03.06.2016. Grant Date: 14.11.2017.
WPS27	Egnor, Daniel; Reid, Elizabeth Hamon. Category Suggestions Relating to A Search. Applicants: Google Inc. Agents: Harrity & Harrity, LLP. US7523099. Application Date: 30.12.2004. Grant Date: 21.04.2009.
WPS28	Gollapudi, Sreenivas. Search Result Diversification. Applicants: Microsoft Corporation. Agents: Shook, Hardy & Bacon L.L.P. US20120089588. Ap plication Date: 19.12.2011. Grant Date: 21.08.2012.
WPS29	Tong, Simon; Pearson, Mark; Brin, Sergey. Methods And Systems for Improving a Search Ranking Using Related Queries. Applicants: Google Inc et al. Agents: Fish & Richardson P.C. US20120011117. Application Date: 18.08.2011. Grant Date: 19.02.2013.
WPS30	Koningstein, Ross. Ranking Documents. Applicants: Google Inc. Agents: Harrity & Harrity, LLP. US08244722. Application Date: 05.01.2010. Grant Date: 14.08.2012.
WPS31	Buron, Florian Michel et al. Viewport-relative Scoring for Location Search Queries. Applicants: Google Inc. Agents: Morgan, Lewis & Bockius LLP. AU2007333558. Application Date: 11.12.2007.
WPS32	Adams, Henele I.; Kim, Hyung-Jin. Modifying Ranking Data Based on Document Changes. Applicants: Google Inc. et al. Agents: Fish & Richardson P.C. US09002867. Application Date: 30.12.2010. Grant Date: 07.04.2015.
WPS33	Acharya, Anurag. Method for Scoring Document. Applicants: Google Inc. JP2007128547. Application Date: 09.01.2007. Grant Date: 08.10.2010.

Code	WIPO's Patentscope
WPS34	Mishne, Gilad; Lin, Jimmy. Search Relevance Using Messages of a Messaging Platform. Applicants: Twitter, Inc. Agents: Steven M. Greenberg, Esq.; CRGO Law. US09836461. Application Date: 05.03.2013. Grant Date: 05.12.2017.
WPS35	Subasic, Pero et al. Systems and Methods for Online User Profiling And Segmentation. Applicants: AOL Advertising Inc. Agents: Finnegan, Henderson, Farabow, Garrett & Dunner, LLP. US08504411. Application Date: 14.09.2009. Grant Date: 06.08.2013.
WPS36	Fontoura, Marcus et al. Indexing System. Applicants: Google Inc. Agents: Brake Hughes Bellermann LLP. US09483568. Application Date: 16.12.2013. Grant Date: 01.11.2016.
WPS37	Vuppala, Raj et al. Customizable Ranking of Search Engine Results in Multitenant Architecture. Applicants: Ariba, Inc. Agents: Schwegman Lundberg & Woessner, P.A. US20170024390. Application Date: 22.07.2015. Grant Date: 13.11.2018.

Table 15. Patent Document List from WIPO's Patentscope.

Code	WIPO's Patentscope
WPS38	Shmiel, Tomer et al. Automatic Query Pattern Generation. Applicants: Google Inc.; Google LLC. Agents: Brake Hughes Bellermann LLP. US20170039267. Application Date: 03.08.2016. Grant Date: 05.11.2019.
WPS39	Leal, Joa˜o et al. Systems And Methods for Document Processing Using Machine Learning. Applicants: Novabase Business Solutions, S.A. US20180300315. Application Date: 11.04.2018.
WPS40	Rappaport, Alain Thierry; Adamson, Daniel. Domain-based Ranking in Document Search. Applicants: Microsoft Technology Licensing, LLC. US20180046716. Application Date: 24.10.2017.
WPS41	Morris, Stephen Timothy. Method For Determining Document Relevance. Applicants: Stephen Timothy Morris. GB2472250. Application Date: 31.07.2009.
WPS42	Dean, Jeffrey et al. Document Scoring Based on Query Analysis. Applicants: Google LLC. AU2007200526. Application Date: 07.02.2007.
WPS43	Dean, Jeffrey et al. Document Scoring Based on Query Analysis. Applicants: Google Inc. et al. US20120209838. Application Date: 24.04.2012. Grant Date: 28.01.2014.
WPS44	Epstein, Samuel S. Methods and Apparatuses for Searching Content. Applicants: Samuel S. Epstein. US20180300410. Application Date: 14.06.2018.
WPS45	Blair-Goldensohn, Sasha et al. Sentiment detection as a ranking signal for reviewable entities. Applicants; Google Inc.; Google LLC.

Code	WIPO's Patentscope
	Agents: Middleton Reutlinger. US10394830. Application Date: 11.03.2016. Grant Date: 27.08.2019.
WPS46	Lawrence, Steve. Document scoring based on traffic associated with a document. Applicants: Google Inc. Agents: Harrity & Harrity, LLP. US20070088693. Application Date: 30.11.2006. Grant Date: 20.11.2012.
WPS47	Buron, Florian Michel et al. Universal scores for location search queries. Applicants: Google Inc. Agents: Morgan, Lewis & Bockius LLP. US08463774. Application Date: 15.07.2008. Grant Date: 11.06.2013.
WPS48	Swapnil, Sanjay Kulkarni. Ranking Entity Based Search Results Based on Implicit User Interactions. Applicants: salesforce.com, inc. US20190205472. Application Date: 28.12.2017.
WPS49	Rajaraman, Anand. Double iterative flavored rank. Applicants: Wal-Mart Stores, Inc. et al. Agents: Bryan Cave LLP. US09286387. Application Date: 22.06.2005. Grant Date: 15.03.2016.
WPS50	Srinivasan, Seshadri et al. Delivering search results. Applicants: Wal-Mart Stores, Inc. et al. Agents: David R. Stevens; Stevens Law Group. US08849830. Application Date: 13.10.2006. Grant Date: 30.09.2014.

Table 16. Patent Document List from WIPO's Patentscope.

Code	WIPO's Patentscope
WPS51	Legrand, Diego Guy M. et al. Bayesian visual interactive search. Applicants: Sentient Technologies (Barbados) Limited. Agents: Haynes Beffel & Wolfeld LLP; Warren S. Wolfeld; Andrew L. Dunlap. US20170091319. Application Date: 09.12.2016. Grant Date: 16.10.2018.
WPS52	Roitman, Haggai et al. Enhanced Query Performance Prediction for Information Retrieval Systems. Applicants: International Business Machines Corporation. US20200210438. Application Date: 31.12.2018. Publication Date: 02.07.2020.
WPS53	Chang, Keng-hao et al. Dynamic tensor attention for information retrieval scoring. Applicants: Microsoft Technology Licensing, LLC. Agents: Merchant & Gould. US20180165288. Application Date: 14.12.2016. Grant Date: 29.10.2019.
WPS54	Henry, Daniel J.; Bascobert Michael R. System and Method for Search. Applicants: AccuPatent, Inc. Agents: Daniel J. Henry. US20090228777. Appli cation Date: 19.02.2009.
WPS55	Mason, Hilary; Levy, Todd. System And Method for Relevance Scoring Of A Digital Resource. Applicants: Bitly, Inc. US20200104946. Application Date: 03.12.2019.

Code	WIPO's Patentscope
WPS56	Barnett, Russell Clark. Metasearch Technique That Ranks Documents Obtained from Multiple Collections. Applicants: NextPage, Inc. Agents: Burns, Doane, Swecker & Mathis, L.L.P. US20020198869. Application Date: 20.06.2001. Publication Date: 26.12.2002.
WPS57	Howard, Jr. Albert R. et al. System and Method for Ranking The Relevance Of Documents Retrieved by a Query. Applicants: Apple, Inc. Agents: Fenwick & West LLP. US7283997. Application Date: 14.05.2003. Grant Date: 16.10.2007.
WPS58	Victor Jr, David Uy. System And Method for Categorically Scoring Electronic Documents. Applicants: David Uy Victor Jr et al. US20160283489. Application Date: 08.06.2016.
WPS59	Victor Jr, David Uy. System And Method for Displaying a Subjective Score with Electronic Documents. Applicants: David Uy Victor Jr et al. US20160170984. Application Date: 23.02.2016.
WPS60	Aphinyanaphongs, Yin; Aliferis, Constantin. Content And Quality Assessment Method and Apparatus for Quality Searching. Applicants: Yin Aphinyanaphongs; Constantin Aliferis. Agents: Laurence Weinberger. WO2010021723. International Filing Date: 20.08.2009.
WPS61	Annau, Thomas M. et al. Selection of Documents to Place in Search Index. Applicants: Google Inc. Agents: Harrity & Harrity, LLP. US08255386. Application Date: 30.01.2008. Grant Date: 28.08.2012.

Table 17. Patent Document List from WIPO's Patentscope.

Code	WIPO's Patentscope
WPS62	Carbune, Victor; Anders, Pedro Gonnet. Contextual Estimation of Link Information Gain. Applicants: Google LLC. Agents: Brantley Shumaker et al. WO2020081082. International Filing Date: 18.10.2018.
WPS63	Grois, Dan. Method for Enabling a User to Vote for a Document Stored Within a Database. Applicants: Dan Grois. Agents: Dan Grois. US20080250105. Application Date: 04.06.2008.
WPS64	Smyros, Athena Ann; Smyros, Constantine John. Systems and Method for Searching an Index. Applicants: Athena Ann Smyros; Constantine John Smyros. Agents: Slater Matsil, LLP. US20160012056. Application Date: 23.02.2015. Grant Date: 21.03.2017.

Table 18. Patent Document List from INPI's Patent Database.

Code	INPI's Patent Database
IPD1	Vicario, Enrico et al. Agendador de Trabalhos para Sistema Eletromecaˆnico de Ana´lises Biolo´gicas. Applicants: Bio Merieux (FR); Universita` di Firenze. Agents: Andre Luiz Souza Alvarez. BR 11 2014 016329 4 A8. Application Date: 13.12.2012.
IPD2	Bice, Anthony Nino et al. Enriquecimento de Respostas de Consulta de Banco de Dados com o Uso de Dados de Fontes de Dados Externos. Applicants: Microsoft Technology Licensing, LLC (US). Agents: Dannemann, Siemsen, Bigler & Ipanema Moreira. BR 11 2013 031948 8 A2. Application Date: 05.06.2012.
IPD3	Gao, Yi et al. Me´todo e Dispositivo para Prover Informac¸a˜o de Contato. Applicants: Xiaomi Inc. (CN). Agents: Kasznar Leonardos Propriedade Intelectual. BR 11 2015 019848 1 A2. Application Date: 30.04.2015.
IPD4	Schreiner, Wilson Douglas. Sistema e Me´todo de Localizac¸a˜o e de Gerenciamento de Informac¸o˜es para Localizac¸a˜o de Pessoas Atrave´s de Dispositivo Sem Fio. Applicants: Wilson Douglas Schreiner (BR/AM). Agents: Toledo Correˆa Marcas e Patentes S/C Ltda. PI 0802762-5 A2. Application Date: 05.05.2008.
IPD5	Jones, Scott, A.; Cooper, Thomas, E. Ferramenta Automatizada para Minerac¸a˜o com Ajuda Humana e Captura de Resultados Precisos. Applicants: Chacha Search, Inc. (US). Agents: Guerra Adv. PI 0707294-5 A2. Application Date: 12.01.2007.

Table 19. Patent Document List from INPI's Patent Database.

Code	INPI's Patent Database
IPD6	Carlock, Thomas V. Aprimoramento de um Questionamento Para uma Busca de um Banco de Dados. Applicants: The Dun and Bradstreet Corporation (US) Agents: Luiz Leonardos & Cia - Propriedade Intelectual. BR 11 2012 028553 0 A2. Application Date: 06.05.2011.
IPD7	Hartwig, Charles D. et al. Sistemas e Me´todos para Adquirir, Analisar e Explorar Dados e Informac¸a˜o. Applicants: Veridex, LLC (US). Agents: Dannemann, Siemsen, Bigler & Ipanema Moreira. PI 0706683-0 A2. Application Date: 19.01.2007.
IPD8	Williams, Shane F.; Ball, Steven J. Sistema Gerenciador de Arquivos Empregando Representac¸a˜o de Dados Baseada em Linha de Tempo. Applicants: Microsoft Corporation (US). Agents: Nellie Anne Daniel Shores. PI 0506009-5 A2. Application Date: 21.11.2005.

IPD9	De Paiva, Sebastia˜o Leandro Morais. Sistema Integrado de Busca e Compra de Produtos por Meio Virtual. Applicants: Sebastia˜o Leandro Morais de Paiva (BR/MG). Agents: Vinícius Silva de Oliveira. BR 10 2017 000526 7 A2. Application Date: 10.01.2017.
IPD10	de Aguiar, Rafael Alexandre; Moraes, Vanessa De Oliveira. Central de Controle Integrada A Aplicativos De Tecnologias Mo´veis E Computador, para Buscas por Servic¸os Atrave´s de Geolocalizac¸a˜o, Com Ferramentas de Contato e Propaganda. Applicants: Rafael Alexandre De Aguiar (BR/SP); Vanessa de Olithemveira Moraes (BR/SP). Agents: Modal Marcas e Patentes Ltda. BR 10 2017 018178 2 A2. Application Date: 24.08.2017.
IPD11	Segawa, Leonardo Hitoshi. Site com Ferramenta para Criac¸a˜o e Edic¸a˜o de Anu´ncios Imobilia´rios em Formato de V´ıdeo Digital, Gerenciamento e Exposic¸a˜o, para Sistema de Busca. Applicants: NT Restaurac¸a˜o Digital, Finalizac¸a˜o e Desenvolvimento Audio Visual LTDA ME (BR/SP). Agents: Modal Marcas E Patentes Ltda. BR 10 2018 000113 2 A2. Application Date: 03.01.2018.