Iterative Integration of Visual Insights during Patent Search and Analysis

Steffen Koch Harald Bosch Mark Giereth Thomas Ertl

Visualization and Interactive Systems Group*
Universität Stuttgart

Abstract

Patents are an important economic factor in today's globalized markets. Therefore, the analysis of patent information has become an inevitable task for a variety of interest groups. The retrieval of relevant patent information is an integral part of almost every patent analysis scenario. Unfortunately, the complexity of patent material inhibits a straightforward retrieval of all relevant patent documents and leads to iterative, time-consuming approaches in practice. With 'PatViz', a new system for interactive analysis of patent information has been developed to leverage iterative query refinement. PatViz supports users in building complex queries visually and in exploring patent result sets interactively. Thereby, the visual query module introduces an abstraction layer that provides uniform access to different retrieval systems and relieves users of the burden to learn different complex query languages. By establishing an integrated environment it allows for interactive reintegration of insights gained from visual result set exploration into the visual query representation. We expect that the approach we have taken is also suitable to improve iterative query refinement in other Visual Analytics systems.

Keywords: Patent retrieval, information visualization, visual analytics, multiple coordinated views

Index Terms: H.5.2 [Information Interfaces and Presentation (e.g.HCI)]: User Interfaces—Graphical user interfaces (GUI) H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Search process

1 Introduction

Patent analysis has become an inevitable task for a large variety of stakeholders, aiming at different objectives, in today's globalized markets. At the same time the amount of patent applications increases rapidly. According to World Intellectual Property Organization (WIPO) [9] statistics there have been 1.76 million new patent filings in 2006, 727,000 patents have been granted and 6.1 million were in force, worldwide.

Databases accessible through the esp@cenet[1] service of the European Patent Office contain more than 60 million patent documents. Patents are not only a concern for large companies but also for small and medium-sized enterprises, who are likely not to keep up their own specialized patent departments and therefore depend on external service providers. Failure in performing a patent analysis thoroughly can result in a high risk of litigation and probably have severe economic consequences. Even if a company does not intend to apply for patents, the patent landscape of the domain(s) a company is involved in has to be tracked closely and publishing inventions might be reasonable to prevent others from obtaining industrial property rights for them. Monitoring of competitors, trend recognition, technology assessment, freedom to operate analysis, and objecting to infringing/trivial patents are other typical tasks in business life.

Apart from intellectual property specialists who are involved with the patent strategies of a company and reviewers from patent offices, many other parties are interested in patent information. These include experts from the finance sector, patent lawyers, scientists and many more. As a consequence, the need to analyze patent information is high.

Unfortunately, not only the rapidly increasing amount of new patent applications and the already available mass of patent information makes patent analysis a tedious task, but also the complexity of available patent material hinders straightforward access to the information needed. For obvious reasons applicants are trying to produce patent applications that still follow the rules of patentability, but they also aim to phrase them as widely as possible to achieve a maximum of coverage for their patents. Occasionally, there also seem to be tendencies to obfuscate patent texts intentionally, e.g. by using terms that are not typical for the corresponding technical field, probably in order to prevent competitors from obtaining easy access to these patents. Because patent applications can address a large variety of technical fields, even without obfuscation the language used within these sectors might differ greatly due to terminology and phrases that are commonly used within one of these fields. Furthermore, some patent applications are multi-lingual, others are only accessible in the language of the country where they have been applied for. Especially the claims section of patent documents makes use of legalese, which further exacerbates their readability and as a consequence reduces their retrievability for inexperienced patent searchers. In addition to the difficulties described above, all problems common to the retrieval of natural language texts (e.g. ambiguities) further increase the complexity.

As part of the EC-project 'PATExpert' [21], a new interface for advanced patent navigation and visualization, called 'PatViz', has been developed. In PatViz, special attention has been paid to overcoming the problems of patent information overload, to simplify patent search and analysis, as well as to lower the entry barrier for inventors that are not patent experts. Therefore, one of the primary goals was the simplification of iterative patent retrieval cycles by utilizing direct interaction on the visual query representation and result set views to leverage insight integration into subsequent refinement cycles. It is expected that such an approach can be utilized within all Visual Analytics systems that incorporate Information Retrieval facilities for unstructured data and where completeness of the findings is of high importance.

The main body of this paper is organized as follows: After an overview of related work in the field of patent analysis, further state of the art will be referred to in the description of the different visualization facilities. Section 3 provides a closer look at the patent analysis process. Section 4 introduces the PatViz system for visual patent analysis. Additionally, the system's capabilities will be exemplified with a detailed description of a typical patent search use case in Section 5. Subsequently, results of the user evaluation will be presented in 6, followed by the concluding Section 7.

*e-mail: {kochsn, bosch, giereth, ertl}@vis.uni-stuttgart.de

[1] Web based service interface for patent search of the European Patent Office, http://ep.espacenet.com

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA
978-1-4244-5283-5/09/$25.00 ©2009 IEEE

2 Related Work

Patent information is complex and potentially ambiguous; its analysis therefore still requires human effort to allow high quality evaluation. The intelligent combination of interactive visualization and retrieval techniques can support users in acquiring relevant information from the heterogeneous, high dimensional, bulky, often ambiguous and conflicting data, thus helping users to gain new insights. Patents contain nearly all data types enumerated by Shneiderman [15] that play an important role in information visualization. According to Thomas and Cook [20], an application for visual analytics should therefore utilize the capabilities of human visual perception, and provide suitable visualizations and interaction techniques to query, explore, and explain large amounts of information. This is also evident for applications in the field of patent analysis. In the following paragraphs previous work in the field of patent analysis as well as related research and visual analysis systems will be described briefly.

Classic information visualization methods for representing hierarchically structured, network, temporal and spatial data are also important within the patent domain. One area where network visualizations can be applied is citation analysis [16] and patent citation analysis (see, in particular, [8, 11]). Commercial products, such as Thomson Aureka[2] or Delphion Citation Link[3] provide citation link visualizations. Forward as well as backward citation information explicitly relates patent documents to each other and is therefore a valuable resource that could be exploited to increase recall in patent search. An example for hierarchical structures within patent metadata is the International Patent Classification[4] (IPC) or the United States Patent and Trademark Office (USPTO) patent classification, which is used in [10] to visualize the evolution of patent spaces by displaying sequences of Treemaps [13].

While these individual visualizations are not particularly suited for depicting multidimensional data on their own, they nevertheless offer an important means for highlighting special aspects of patent data. They can be integrated into systems handling multidimensional data by using multiple coordinated views [12] in conjunction with brushing and linking or by using them within focus & context techniques [7].

Systems for the evaluation of structured data such as Polaris [19] enhance analysis tasks by providing techniques for direct manipulation of data views. WireVis [5], for example, allows online reclustering of large amounts of transaction records to search for suspicious patterns.

While those systems provide feasible approaches for relational data, e.g. from data warehouses, the situation in patent analysis differs from such a scenario due to the incorporation of text (and image) retrieval leading to unsharp and uncertain results. Good examples for systems that address the analysis of text document collections are Jigsaw [18] and IN-SPIRE [23]. Contrary to our system both approaches focus on medium sized and locally available collections. Jigsaw employs entity co-occurrence analysis in short reports to guide the user on which report to read next. IN-SPIRE projects the documents into a two dimensional view and allows for comparing groups that relate to combinations of search queries.

The iterative cycle of (i) query formulation, (ii) visual representation of result sets, and (iii) examination of the results is an important aspect of patent search and analysis. Current publicly available systems for patent search such as esp@cenet or DEPATISnet[5] provide either form-based or textual interfaces for query formulation and represent the results as lists and/or text fragments. Some of the commercially available systems like Matheo Analyzer[6] and ANAVIST[7] provide visualization of result sets either based on patent metadata or patent texts. Another advanced interface that provides better support for iterative query refinement, but only by the means of a textual interface, is Questel's QPAT[8]. Other research that is significant in the context of this paper includes methods that build queries visually [1, 17, 3] - an overview can be found in [4]. As opposed to these query formulation approaches, our system adds the possibility to integrate multiple query languages and ways to interconnect query and result set view through insight integration.

3 The Patent Analysis Process

Patents are conferred to applicants for a maximum time span of twenty years. While in force, they protect the applicants' invention, thereby granting them the exclusive right to decide who is allowed to make commercial use of the invention and who is not. By applying for a patent, applicants automatically accept the publication of their patents. This means that other parties can make use of the information that is contained within the patent document. As a consequence, the need to analyze patent information is high.

3.1 Patent searching

While the motivation of different groups of patent information users may differ greatly, searching for relevant patent information is part of almost every patent analysis task. Compared to web search, retrieving patent information is a lot more demanding, and not only due to the problems described above. While a user's need in web search is generally satisfied by a few precise hits, patent analysis must usually take into account all relevant patent documents. It is very difficult to query patent data without missing relevant information and/or without getting overwhelmed by very large result sets that cannot be handled efficiently by users any more. Much expertise and time is required to find a complete patent subset comprising all the relevant information without containing too much noise.

All the mentioned characteristics lead to a process for patent analysis that is common in the domain. Figure 1 depicts this process, which consists of three stages, namely *patent retrieval*, *patent result set analysis*, and *patent detail analysis*. Beginning with an initial search query that describes the object of investigation roughly, the examiner iteratively refines the query. Hereby, obviously irrelevant clusters of patent documents are excluded based on aspects that were found in the documents returned from the search engine. On the other hand, if the examiner finds new features that also describe the object of investigation, e.g. synonyms of already used keywords, the query is refined to widen the result set. These refinements are repeated until the examiner is confident that all relevant patent documents are contained within the result set and that it does not comprise more documents than can be handled efficiently. It is not unusual to work with queries that are constructed from 30-40 preliminary queries.

The process described above explains why Boolean retrieval systems are highly popular in the patent domain. The examiner relies on a system that includes and excludes features from the result set precisely as the query describes. Otherwise queries of that size could not be handled and the relevance of the result set could suffer. But not only the natural language text content of patents hinders their retrievability; it is also the large amount of different types of data that makes patent information analysis a challenging task.

3.2 Patent Data

Patent information comprises nominal data (e.g. from legal entities like inventors' and applicants' names), structured data, such as addresses from these legal entities, location and time-based data (e.g.

[2] http://thomsonreuters.com/content/PDF/scientific/corp /AurekaFactSheet.pdf

[3] http://www.delphion.com/products/research/products-citelink

[4] http://www.wipo.int/classifications/ipc/en/

[5] http://www.dpma.de/service/depatisnet.html

[6] http://www.matheo-analyzer.com/

[7] http://www.stn-international.de/stninterfaces/stnanavist/stn_anavist.html

[8] http://www.qpat.com/index.htm

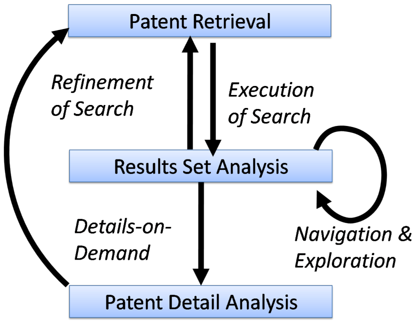


Figure 1: The patent analysis loop

contracting states, legal events), hierarchical data like classificatory information, relational data that builds network structures like in the case of patent family[9] information or citation relations, and of course unstructured data such as patent text and drawings. An important kind of patent metadata is classification information that, according to the applied system (e.g. IPC), classifies patent documents, for example by the technical field(s) they were applied for. Depending on the user's analytical task, different aspects from this pool of data might be of interest, either to start the patent retrieval with, or to drill down to certain aspects within the relevant set of already retrieved patent documents. A much more comprehensive overview of data stored during the life-cycle of a patent application can be derived from the standards for electronic exchange of patent document information, which have been defined by the WIPO standard ST.36[10].

4 PatViz

As mentioned above, support for iterative refinement of search requests should be an essential objective of patent search software. Thereby, query modification typically aims at either narrowing the request to remove noise from the results, or at widening the search because the user wants to find additional relevant patents. In either case, the insight that generates a user's desire for refinement is normally perceived through the display of the previous result set data. Hence, patent search software needs an efficient feedback loop to transfer insights from result set exploration to the query formulation. This differs from approaches suggested by Shneiderman in [2] and realized in systems like Polaris [19]. The main dissimilarity is that the underlying backend systems do not guarantee the completeness of a request's result. Because completeness of results is of high priority in patent search, users are forced to build trust in their retrieved results by carrying out the iterative query process described above.

PatViz was built as a graphical front-end for a set of different search engines and patent document analysis services developed in PATExpert. However, the patent domains accessible by PATExpert's backend systems were restricted[11] to the IPC main classes 'optical recording' and 'machine tools'. The essential components of PatViz comprise a querying system, a multitude of visual result set representations and the linkage between them. All these components are bundled in a desktop-like application that handles the data management and event propagation between the components.

The remainder of this section mirrors the structure of the PatViz system itself and introduces its essential components.

4.1 Query System

Querying the retrieval system is the initial task that has to be performed when working with a patent information system. This is true for almost every use case independent of the analysts' concrete goals. The central idea of our approach is the tight linkage between query reformulation and result representation. To understand this linkage one has to understand the querying system of PATExpert.

As described in Section 3, queries in the patent domain tend to get complex and large. In PATExpert, it is possible to integrate different search facilities in one query, adding even more potential for complex queries. In its current state PATExpert offers the following search facilities [6]: *full text search*, *metadata search*, *image similarity search*, *semantic search*, and *document similarity search*, whereby the latter represents a special case of full text search.

The full text search engine provides conventional keyword search in our patent analysis systems. Patent full texts as well as all metadata are stored within a relational database. The image similarity search is accomplished by a system based on a vector space model. Thereby, feature vectors are computed from the images through several preprocessing steps. The semantic information extracted from the patent documents is stored in a W3C semantic web format, which is accessible through a semantic repository.

Each query subsystem has its own formal query language. To facilitate the usage of all query subsystems in one coherent interface, a possibility to integrate them as well as their query languages had to be developed first. The combination of different search expressions from different search facilities is possible through a Boolean integration language.

4.1.1 Boolean Integration of Search Facilities

Providing different search engines that can be combined through a Boolean integration language allows for stating complex and powerful queries, but also makes query creation a sophisticated task for the user. To compensate the complexity of the new, combined query language a visual query editor has been developed that is directly linked to a conventional textual interface. As a requirement, the editor had to provide a clear view of the logical structure of the whole query and a form-based way to create search expressions for each of the different facilities. The result of this integration can be seen in Figure 2.

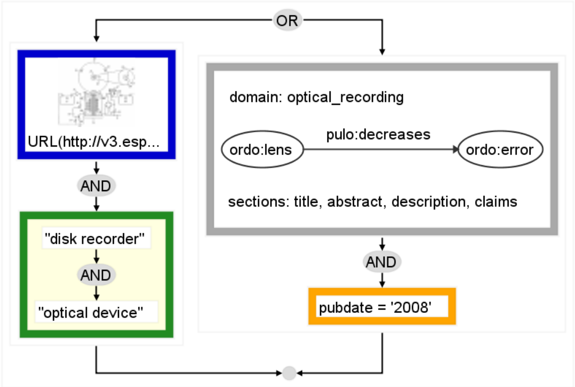


Figure 2: The graphical representation of combined search expressions for different retrieval facilities: Image similarity search (blue), semantic search (grey), keyword search (green), and metadata search (orange).

[9] There exist several distinct definitions of 'patent families'. However, they can be interpreted at least as hierarchical, typically as a network-based data structure of interlinked patent applications.

[10] http://www.wipo.int/standards/en/pdf/03-36-01.pdf

[11] This restriction was necessary to perform the natural language processing on the patent material, which was a prerequisite for the fulfillment of other scientific objectives in PATExpert.

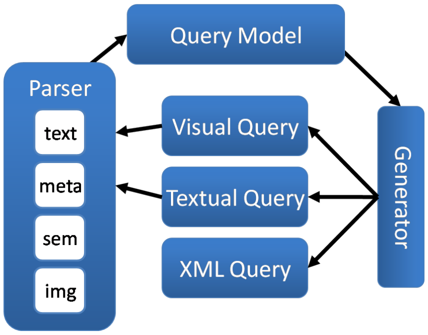
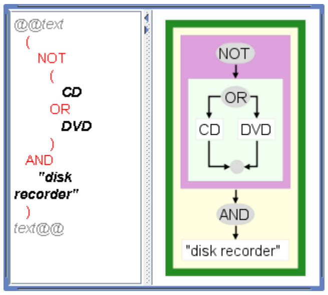


Figure 3: Different query representations can be processed by the hierarchical parser and are transformed into a general query model, which again allows for generating the other representations.

Technically this is realized by a hierarchical parser/generator module as depicted in Figure 3. The module is capable of parsing/accepting the textual as well as the visual representation of the combined query language. If expressions/visual constructs of a specific sublanguage are encountered, they are forwarded to the corresponding subparser. If the query was syntactically correct, which is not guaranteed in the case of textual query creation, the other representation is automatically updated. If a query is going to be sent to the search system again, an XML representation of the query is generated to be encapsulated in a web service request. The interpretation of this XML request, the decomposition into the different sublanguages, querying each search service, merging of the results, elimination of duplicate results, and their delivery back to the visualization module is the job of PATExpert's merger service as sketched in Figure 5. Due to the hierarchical parser/generator concept the query system can also be adapted to other domains or extended by additional search facilities.

To create an appropriate metaphor for the Boolean integration language, we decided to use one that is closely related to the very common Syntax Diagrams [22]. Therefore, our approach uses node-link diagrams with an orthogonalized circuit-like graph layout as displayed in the right half of Figure 4. The set of operators for the Boolean integration language is limited to 'AND', 'OR' and 'NOT'.



```
@@text
 (
  NOT
   (
    CD
   OR
    DVD
   )
 AND
   "disk
recorder"
 )
text@@
```

Figure 4: Orthogonalized layout of a text query with Boolean operators in text and graphical form.

Links describe a combination of these constraints correlating directly with the binary operators 'AND' and 'OR' inside the visualization. A sequential link between two nodes always expresses the 'AND' relation and has the semantical meaning that both constraints represented by the connected nodes have to be fulfilled by a patent to pass the filter function. A branching link on the other hand represents an alternative ('OR') and has its semantic equivalent in a conjunction of filtered results of every branch that belongs to the same junction. The 'NOT' operator is represented by a box, which encloses the negated constraint. The boxes around each operator identify the scope of a 'NOT' operator very clearly, as can be seen in red in Figure 4, too. Users can identify which terms are part of which boolean operator, where the area of control of an operator ends, and they can spot the positions where they might want to alter the query.

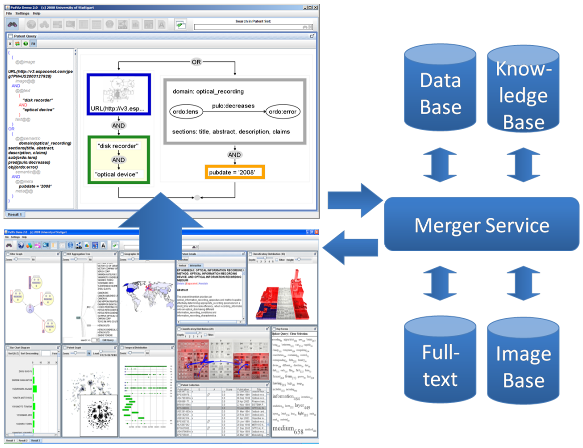


Figure 5: The flow of communication within the querying system. The result query builder (upper left) is capable of incorporating user triggered update information from the result set view (lower left) directly. The result set view is updated indirectly after re-querying the merger service, which is connected to the different search facilities (right).

4.1.2 Creation of Search Statements

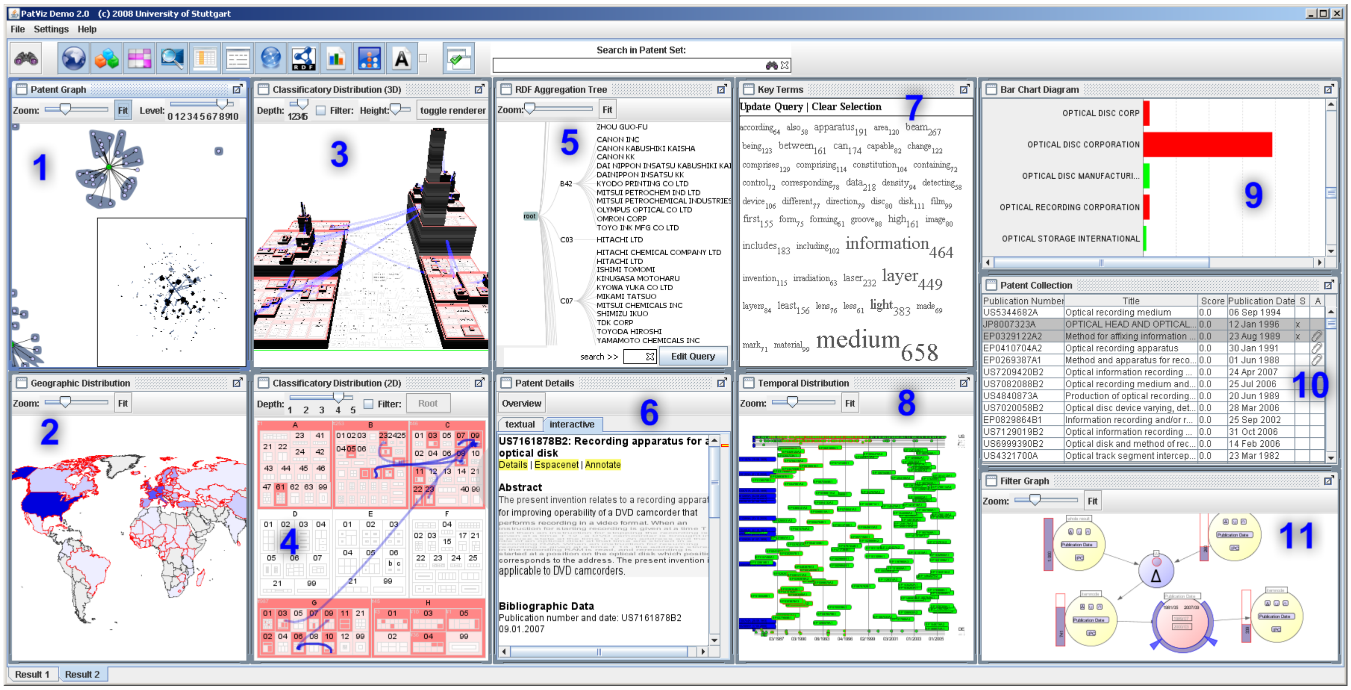
The creation of query terms is supported by forms corresponding to the specialized sub-queries. This ensures that only syntactically valid search terms can be created and it frees the user from the cognitive task of remembering possible filter operations and values by presenting them.

Further interaction functionality allows to zoom and pan the graphical representation of the constructed query. Within the hierarchical graph representation, nodes can be zoomed in while other nodes are put into the background and complex nodes can be collapsed/expanded to further enhance comprehensibility of the query graph.

A string containing the query text is displayed in addition to the graphical representation. It also accentuates the logical structure of the query by reformatting the input with line breaks and indentation like in the example in the left half of Figure 4. On the one hand, this ensures that expert users who are familiar with the query language of the different search engines can still enter queries directly. On the other hand, having both representations available can help inexperienced users to learn the query language vocabulary.

4.2 Result Set Exploration

The second group of components that is relevant to our interactive query refinement task is the representation of the query's result set. PatViz provides ten different views of the result set, which are shown in Figure 6. For their integration into the PatViz environment, all views must provide interfaces for basic brushing and linking operations. That means they must be able to understand and create selection events. Every brushing operation in one view results in the selection of a subset of patent documents. This subset is encapsulated in a selection event and distributed through the PatViz environment. When other views receive such an event they have to display the selection within their view appropriately. For



Publication Number	Title	Score	Publication Date	S	A
US5344682A	Optical recording medium	0.0	06 Sep 1994		
JP8007323A	OPTICAL HEAD AND OPTICAL...	0.0	12 Jan 1996	x	
EP0329122A2	Method for affixing information ...	0.0	23 Aug 1989	x	📎
EP0410704A2	Optical recording apparatus	0.0	30 Jan 1991		📎
EP0269387A1	Method and apparatus for reco...	0.0	01 Jun 1988		📎
US7209420B2	Optical information recording ...	0.0	24 Apr 2007		
US7082088B2	Optical recording medium and...	0.0	25 Jul 2006		
US4840873A	Production of optical recording...	0.0	20 Jun 1989		
US7020058B2	Optical disc device varying, det...	0.0	28 Mar 2006		
EP0829864B1	Information recording and/or r...	0.0	25 Sep 2002		
US7129019B2	Optical information recording ...	0.0	31 Oct 2006		
US6999390B2	Optical disk and method of rec...	0.0	14 Feb 2006		
US4321700A	Optical track segment intercep...	0.0	23 Mar 1982		

Overview

textual | interactive

US7161878B2: Recording apparatus for a optical disk

Details | Espacenet | Annotate

Abstract

The present invention relates to a recording apparat
for improving operability of a DVD camcorder that
performs recording in a video format. When an
instruction for starting recording is given at a time T
[illegible]
started at a position on the optical disk which positio
corresponds to the address. The present invention i
applicable to DVD camcorders.

Bibliographic Data

Publication number and date: US7161878B2
09.01.2007

Figure 6: All available result set views in the PatViz desktop. These are from top to bottom and from left to right: 1. *Patent Graph* - a configurable graph view that can show various connections between entities of the result set; 2. *World Map* - a distribution of the patent documents over the filing countries; 3. *3D IPC Treemap* - a distribution of the patent documents over a classification schema shown in a 3D Treemap; 4. the same in 2D; 5. *Aggregation Tree* - a tree view that can aggregate the result set by an adjustable hierarchy; 6. *Text View* - a viewer for patent document texts that can overlay results of the linguistic analysis and allow for intra-document navigation; 7. *Term Cloud* - a cloud of words that refers to the most frequent terms; 8. *Geo-Timeline* - a scatterplot of the filing date and filing country of patent documents; 9. *Bar Charts* - a simple bar chart aggregation of the set by one choosable metadata field; 10. *Table* - a table containing the most important data of the patent documents like number, title, and applicant; 11. *Selection Management* - a graph based tool to store, combine and adjust selections.

example, the world map varies the saturation and color of the countries according to the number of selected documents filed within the country.

However, simply selecting a subset of documents is not enough to be able to reflect interesting subset definitions to the search query. Therefore the views also need to deliver a description of the filter operation used to create the subset. In the example of the world map, this could be 'filing-country = sweden' if users mark Sweden, or a concatenation of such statements if they select a multitude of countries. With this increased self descriptiveness of selections we can (1) enrich the selection management with interactive adjustments of the underlying filter mechanisms and (2) create appropriate filter definitions for the search query reformulation.

Within result set visualization environments, traditional interaction techniques for selecting specific data subsets from the visualized information space may not be sufficient. Especially if the selection operation comprises separated sub selections in multiple linked views of aggregated data, the selection of a particular data set may be difficult or even impossible. Therefore 'PatViz' contains a graph-based tool for visual selection management allowing for the combination of data subsets by applying set operations on them. This tool provides increased expressiveness over classical approaches by utilizing them as building blocks for more complex extraction strategies.

The tool itself provides a graph view (see view 11 in Figure 6 and Figure 9) that contains nodes serving as interactive widgets. The graph, which can be built in a user-steered process, is directed and comprises three different types of nodes: *content nodes*, *filter nodes*, and *operator nodes*. At the beginning of the interaction there is always a single source content node representing the entire set of patent documents contained in the current result set. Content nodes have a vertical bar attached to them symbolizing the size of the set they represent in relation to the whole set as imported from another view. Additionally the bar is labeled with the exact size of the set. Content nodes can be connected to filter nodes, which constrains one of the set's attributes, in order to restrict a content node's set of documents. The result of the restriction is another content node with the reduced document set. The third type of nodes constitutes set operator nodes. These nodes allow for the combination of different content nodes and thus can have an arbitrary number of incoming edges and one outgoing edge, each again connected to corresponding content nodes.

Boolean combinations of filtering constraints are expressed directly by the graph structure. We allow the usage of explicit set operators, i.e. union and intersection, via additional nodes besides the implicit combination contained in the graph structure itself, i.e. sequences and branches. In contrast to the filter/flow metaphor [14] these operators facilitate the combination of arbitrary sets of data objects without the need to generate multiple instances of a particular filter just to apply it in different combinations. Arbitrary input nodes can be created for our graph simply by performing a 'classic' selection operation in one of the other result set views.

The construction of the graph itself is performed completely by the users. By guiding users when they interact with the graph widgets, the creation of illegal graph configurations is prevented. Operations on the selection and filter component can be performed via direct manipulation. This applies to all types of nodes and their Boolean combination. Different filter nodes are created with respect to the data type of the property that should be constrained. After combining different sets and parameterizing filter nodes, every document subset can be reflected back to the result set visualization by selecting an arbitrary content node.

4.3 Query Refinement through Insight Integration

This subsection emphasizes the importance of the described mechanisms to integrate result set related insights to selection management and query formulation. Here, several levels of insight integration in the PatViz system can be identified.

As stated earlier, reading documents in the patent domain tends to be rather laborious, but they provide a variety of metadata for creating aggregations, relations, and statistics. Thus, it is possible to create a rich set of views on result sets. Without integration the interactions provided by an individual view are restricted to adjustment of view dependent parameters like sorting, filtering, highlighting, zooming, and panning. The user can only gain insights by exploiting the set's metadata, which is related to the view.

The first level of integration is brushing and linking between the views to make connections in the result set visible. E.g. by cross-highlighting, the user can answer questions about the frequent filing countries of the applicant with the highest number of patents in the set. While being a powerful tool, brushing and linking can only show connections between the selection in one view and its representation in the other views.

The second level of integration is the saving and recombining of selections. Multiple views can now be used to define subsets and combine them employing set operators, allowing the user to answer the same type of questions as above but with additional restrictions from other views. E.g: Who is the applicant with the highest number of patent documents published prior to the year 1989 from Spain within my result set? This question could also be formulated as a new query, but this would make the combination of the answer with other subsets of the result set more complicated.

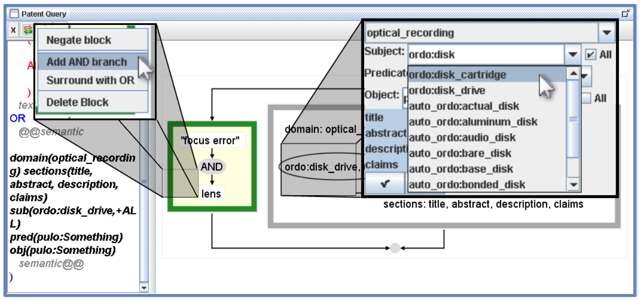
So far, the user did not leave the phase of exploring the result set. While this phase is important for creating insights regarding the problem domain, it opposes the patent domain's need for high 'relevance' values of result sets. Therefore, query widening has to come into play. The third level of integration addresses this need in form of a query refinement by result set interaction. The views are aware of the type of data they are displaying and are capable of providing a search expression based on the user's selection in the corresponding view. The selection management component, in turn, is capable of combining the selections and their attached search term description to complex queries. Finally, the visual query editor allows for the direct incorporation of combined selections, to find more or exclude documents of the specified kind. This aspect cannot be achieved by a single component, but only by the whole system.

Because the selection management tool does not depend on the type of documents in the sets, and because filter options are derived from the underlying data model, the selection management and its insight integration facility can be applied to other application domains without great difficulty.

5 A TYPICAL USE CASE

To illustrate how our system benefits the user, this section describes a short use case from the patent domain. As previously stated, the patent document repository is a great source of technical knowledge. Suppose we are a manufacturer of optical disk drives and are confronted with a technical problem concerning lens focus errors in one of our product lines. Thus we are searching for solutions or available cooperation partners with knowledge in this area.

PatViz starts by showing the visual query editor containing an example text search term. We change the example text directly in the visual query to `focus error` as relevant terms for our problem. Alternatively we could have altered the textual representation on the left hand side of the visual query. Then, we use a context menu to add another mandatory search term, `lens` in our case. The resulting query can be seen as a green box in Figure 7.



```
domain(optical_recordin
g) sections(title,
abstract, description,
claims)
sub(ordo:disk_drive,+AL
L)
pred(pulo:Something)
obj(pulo:Something)
  semantic@@
)
```

Figure 7: The inital query. The text search term has a green border, the semantic search term has a gray border. Zoomed areas illustrate interactions during the creation of the query.

The next step would be to include an expression describing an 'optical disk drive' in the query. However, there are many types and different writings of disk drives. Disk may also be spelt as 'disc' or the patents could refer to special disks like 'DVD' or 'CD'. Therefore, we add a semantic concept search term to unite all these special cases, again by choosing the appropriate item in a context menu, which explicitly defines the location of the new search fragment within the query. Here, a form lets the user specify a relation between two concepts or just a single concept. As we start typing `disk`, an auto completion list appears to show the available semantic concepts as can be seen in right hand side of Figure 7. Our initial query can now be submitted to the search engine, which takes care of separating the queries for the different search facilities and merging the answers in a single result set.

The returned result set contains more than 500 patent documents. A patent searcher would now start reading some documents and titles to find unwanted topics that could be excluded. In PatViz, we take a look at the key terms cloud which summarizes the most frequent terms of the document set. We select a combination of terms (`control`, `circuit`) that seem to refer to electronic solutions to the lens focus problem. Let us assume we are not interested in an electronic solution and want to adjust our search query to exclude these words. The selected terms can easily be transfered to our visual query as a new building block (Figure 8). To exclude patents with this term combination, we simply negate the block via a context menu and re-run the query.

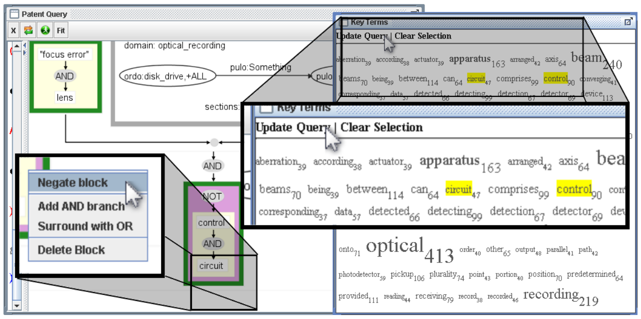


Figure 8: The refined query. After selecting unwanted terms, the query is updated to include them. The new search fragment must be negated to match our intention. Zoomed areas illustrate interactions during this process.

To identify the key player in our problem domain we are looking for the applicant with the highest number of patents within the new result set and the additional constraint that he should be from the same country to allow for easy cooperation. Additionally, patents should be not in force anymore and must therefore be older than 20 years. To achieve this we examine the bar chart of patents per applicant and pick the one with the largest patent amount. We store this selection in the selection management component for later use. After restricting the set to our preferred country, e.g. Spain, we store this selection, too. Then we create an intersection node in the selection management tool and choose the two selections as input. This results in a node with patent documents fulfilling both criteria. The last step is to apply a filter on publication dates to exclude newer patents from the selection. This leaves us with a node containing only 11 patent documents, which can be redisplayed in the other views or read with the document viewer. The described process is depicted in Figure 9

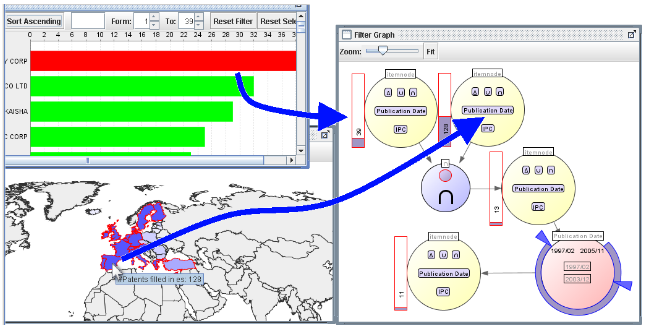


Figure 9: The combination of selections from different views. Nodes are created from the views (blue arrows) and combined using an operator node (shaded blue). After filtering the publication date (red node), we are left with the final result (node at bottom center).

If a closer examination of the remaining documents reveals additional aspects that require a new query iteration to be included, we can import our examination process from the selection management component into the visual query builder. Every action in the result set exploration is then mapped either to a search expression definition or a Boolean combination.

6 Results

To show the suitability of the approaches taken in the PatViz system, two evaluation tasks with two different groups of testers were conducted. The viability of using a visual query representation with respect to its understandability was evaluated through a questionnaire that was sent to the evaluators via email. Fifteen of the evaluators answered it; two of them were patent specialists, the others employees of the computer science department of the Universität Stuttgart.

To receive feedback on our approach for iterative insight integration, a think-aloud evaluation with three patent practitioners knowledgeable in the field of 'optical recording' was conducted.

6.1 Suitability of the visual query builder

As mentioned in 4.1, the visual query system consists of two coordinated views - a text-based and a visual one. All evaluators were asked to answer questions concerning the following aspects: *Feasibility of the chosen visual metaphors*, *Comprehensibility of visual metaphors*, *Recognition of the scopes of Boolean operators*, *Helpfulness of interactive exploration for query understanding*, *Creation of Boolean queries*, and *Composition of complex queries including different search facilities*.

The evaluators disagreed on whether the Boolean AND operator should be represented by a sequential or a branching metaphor (analogous for the OR operator). Nevertheless, none of them had difficulties in correct interpretation of the metaphors. Therefore, there is a strong indication that the visual AND metaphor is nevertheless *feasible*, but this should be verified again in subsequent evaluations. The *comprehensibility* of the provided visual query example has been high. All except one of the testers interpreted the visual example queries correctly. The same holds for the testers' ability to *recognize operator scopes* correctly. Thirteen of the testers considered scope highlighting a useful feature for the exploration of queries. With respect to *creation* of Boolean queries, three testers mentioned that they would prefer a purely textual query interface over a visual one. All others preferred the combined approach which has been applied in PatViz. Twelve of the test persons expressed the opinion that our approach is suitable for the composition of complex queries including the integration of multiple search facilities. Three were undecided. The result of the questionnaire's evaluation suggests that even without using the query tool for direct insight integration the approach already provides an advantage over a purely textual approach.

6.2 Iterative insight integration

The viability of our concept for insight integration into subsequent search/analysis cycles is much more demanding to test. As already discussed, correct interpretation of patent documents requires at least some experience with the technical field under analysis. On the other hand, the employment of patent specialists for this task was a must, in order to be able to judge the suitability of the developed tools. Since it was difficult to find patent specialists knowledgeable in the field of 'optical recording' or 'machine tools', patent practitioners from the consortium were asked to take part in a think-aloud evaluation. The actions of the participants as well as their 'loudly spoken thoughts' were recorded. Naturally, the validity of such a test is limited by the fact that participators were involved in the PATExpert consortium as well as by the relatively small sample for this evaluation. Thus, the results are informally presented.

One frequently expressed comment indicated that most of the patent experts never worked with a system providing interlinked and interactive visual interfaces. While this was also one of the system's properties that was most appreciated by the users, it became clear that such features are very difficult to use without any training. In order to carry out the 'think-aloud' evaluation, the test persons were given access to an online version of the system prior to inviting them for the test itself. Additionally, the evaluators were introduced to brushing and linking within the multiple coordinated views interface and to the meaning and usage of the available views. Subsequently, they were asked to carry out the same analysis tasks they are performing in their daily work.

The following paragraph sketches some identified benefits and flaws of PatViz. Interestingly, all patent users agreed that the visual interface is a valuable means for creating and editing complex queries for different search engines, but some of them were puzzled when they had to use it for the first time. In subsequent discussions it became clear that many, mostly form-based, interfaces for patent search are designed in the same way patent documents are structured. Of course, this is not reflected within an interface that allows for arbitrary combinations of different constraints for search facilities; however, it might be a good starting-point for future enrichment of the query visualization tool with a third view that takes this issue into account. Another observation is that most of the patent experts used views like the tag cloud, the charts, and the world map more frequently than the more sophisticated ones. A probable explanation for this behavior is that users may tend to perform their tasks with tools they are used to. Nevertheless, after a quick introduction, the testers were able to integrate the other views successfully into their analysis. The most significant benefit that has been identified by the test users was the support for iterative refinement of queries and patent sets. Also the synergetic effect of using different views of the same set in parallel were appreciated by the users and the linking and brushing facilities were used extensively after a short period of familiarizing themselves with the system. The testers commented positively on the variability and power of the system resulting from the degrees of freedom in moving back and forth in the stages of the analysis process and between different perspectives within one stage of the process.

To conclude: On the one hand, the flexibility and implicit functionality provided by the developed prototype is difficult to comprehend when users start working with the system without previous instruction. To some extent this problem could be reduced by providing appropriate context-sensitive help systems. On the other hand, a powerful and flexible demonstration prototype that facilitates different patent analysis tasks has been created. This fact has been recognized by the test users and was positively emphasized by them during the test sessions.

Because the development of the visualization module will continue, some of the identified problems will be addressed and reevaluated.

7 Conclusion and Future Work

We presented PatViz, a system for patent search and analysis that exceeds current patent retrieval and analysis systems by providing a flexible 'multiple coordinated views' system. Furthermore, PatViz supports users directly in the iterative refinement of search results, which offers new opportunities for the employment of modern retrieval systems based on vector space models within patent retrieval. This is accomplished by supporting patent searchers in addressing different search facilities within one visual interface, its tight integration of a variety of views presenting different perspectives on patent result sets, and the means to integrate insights interactively from these result sets to improve the query for subsequent refinement of the patent space under analysis.

As part of the DFG Priority Program 1335 'Scalable Visual Analytics' we currently continue our work within the field of visual patent information analysis in cooperation with the Institute for Natural Language Processing of the Universität Stuttgart. The focus will be shifted to the retrieval of textual patent information in order to further increase the reliability of the retrieval results and to enhance users' trust in these new retrieval techniques. Thereby, more emphasis will be placed on Named Entity Recognition in patent texts and the development of new solutions for patent classification problems. PatViz will be used as the basis for our developments in this field of research.

Acknowledgements

The authors wish to thank all partners of the PATExpert consortium. The work presented in this paper was funded by the European Commission in context of the FP6 project 'PatExpert' and is supported by the DFG as part of the Priority Program 1335 'Scalable Visual Analytics'.

References

[1] C. Ahlberg and B. Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *CHI '94: Conference companion on Human factors in computing systems*, pages 313–317, New York, NY, USA, 1994. ACM.

[2] C. Ahlberg, C. Williamson, and B. Shneiderman. Dynamic queries for information exploration: an implementation and evaluation. In *CHI '92: Conference companion on Human factors in computing systems*, pages 619–626, New York, NY, USA, 1992. ACM.

[3] R. Baeza-Yates, G. Navarro, J. Vegas, and P. De La Fuente. A model and a visual query language for structured text. In *String Processing and Information Retrieval: A South American Symposium, 1998. Proceedings*, pages 7–13, Sep 1998.

[4] T. Catarci, M. F. Costabile, S. Levialdi, and C. Batini. Visual query systems for databases: A survey. *Journal of Visual Languages & Computing*, 8(2):215 – 260, 1997.

[5] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. WireVis: Visualization of categorical, time-varying data from financial transactions. In *IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007*, pages 155–162, 30 2007-Nov. 1 2007.

[6] J. Codina, E. Pianta, S. Vrochidis, and S. Papadopoulos. Integration of semantic, metadata and image search engines with a text search engine for patent retrieval. In *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008)*, pages 14–28. CEUR-WS.org, June 2008.

[7] G. W. Furnas. Generalized fisheye views. In *CHI '86: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 16–23, New York, NY, USA, 1986. ACM.

[8] A. Jaffe and M. Trajtenberg. *Patents, Citations & Innovations*. MIT Press, 2002.

[9] M. Khan, R. Lamb, B. Le Feuvre, W. Meredith, C. Calais Regnier, A. Riechel, and H. Zhou. *THE WORLD PATENT REPORT - A STATISTICAL REVIEW, 2008 Edition*. World Intellectual Property Organization, 2008.

[10] D. Kutz. Examining the evolution and distribution of patent classifications. In *Eighth International Conference on Information Visualisation, 2004. IV 2004. Proceedings*, pages 983–988. IEEE Computer Society, July 2004.

[11] L. Reeve, H. Han, and C. Chen. *Information Visualization and the Semantic Web, in Geroimenko, V., Chen, C. (Eds.), Visualizing the Semantic Web*. Springer, 2006.

[12] J. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV '07*, pages 61–71, July 2007.

[13] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, 1992.

[14] B. Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77, Nov 1994.

[15] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages, 1996. Proceedings*, pages 336–343, Sep 1996.

[16] H. G. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

[17] A. Spoerri. Infocrystal. In *CHI '94: Conference companion on Human factors in computing systems*, pages 11–12, New York, NY, USA, 1994. ACM.

[18] J. Stasko, C. Gorg, Z. Liu, and K. Singhal. Jigsaw: Supporting investigative analysis through interactive visualization. In *VAST '07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 131–138, Washington, DC, USA, 2007. IEEE Computer Society.

[19] C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, Jan/Mar 2002.

[20] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[21] L. Wanner, R. Baeza-Yates, S. Brügmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, and V. Zervaki. Towards content-oriented patent document processing. *World Patent Information*, 30(1):21 – 33, 2008.

[22] N. Wirth. *Systematic Programming: An Introduction*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1973.

[23] P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas. In-spire infovis 2004 contest entry. In *IEEE Symposium on Information Visualization*, Oct. 2004.