What kind of semantics and why to use it in patents
Part 1 – Semantic search

Alberto Ciaramella – Marco Ciaramella (IntelliSemantic)

Executive abstract

The term semantics means to add meanings to text.

The world "sematic" entered in patent information circles about ten years ago, as "semantic search", i.e. a new query paradigm, which starts searching from a prototype text.

The interest about this topic, is increasing now, as demonstrated by new presentations on recent patent events, as CEPIUG 2018 and EPOPIC 2018. This topic still attracts some concerns, mostly due to the fact that semantic searches are perceived as black box solutions. This contribution will try to "open the box" of semantic searches, summarizing different features of semantic searches, which can be used for a first assessment of specific solutions. Moreover, from these features we can infer that semantic searches available today are very different together, and can be differentiated into two categories or generations.
Finally we suggest that the preferred use today for semantic searches is to increase the recall after a "traditional" search.

In any case, semantic in patents can be used also in other cases besides searches, as in patent analyses, whilst new technical evolutions are still welcome: these topics will be detailed in the following chapters 2 and 3 of this whole contribution.

Table of Contents

Tables

1. Semantic search

1.1 Framework

The term semantics, a specific branch of natural language processing, means to add meanings to text, since a word could have different meanings, depending from the context.

The world "semantic" entered in patent information circles about ten years ago, mainly as "semantic search" solutions, which used a new query paradigm, i.e. a prototype text, as a patent or also a news: by the way, this is only a specific application of semantics.
This new query paradigm is added to other well established query paradigms.
All query paradigms are summarized in table 1, which distinguishes explicit queries (by query languages or by predefined forms) and implicit queries, based on a prototype text, as summarized in table 1.

Cathegory	Subcathegory	Advantages	Notes
Explicit query	using a query language	More general to formulate	"Some semantic" can be included b using IPCs, CPCs
	using predefined query forms	Simpler to formulate	"Some semantic" can be included b using IPCs, CPCs
Implicit query (example text based)	Citation based	Relies also on third party knowledge, i.e. citations	To be considered for granted patents which include examiners citations
	Text based (or "semantic query")	More general: to be used if citations are not available or weak	More realistic for patent application and for other kind of reference text

Table 1: An overview of patent queries

Explicit queries use text entities and strings, together with metadata, i.e. applicant, dates, technology domains, coded e.g. as IPCs or CPCs. Technology domains add a meaning to text, hence "some semantics" is also implicitly used in these more traditional approaches.

Implicit queries, using a reference text can be further distinguished into citation-based and text-similarity based.

The advantage of citation-based solutions is to include additional knowledge besides the original patent, since citations imply a third party judgements. Citation-based solutions can be further categorized by specific implementation choices, as for example if they manage or not the citation source (applicant, examiner) and how to handle citation patterns (e.g. backward, forward, direct, indirect).

Citation-based solutions are useful when starting from a patent document including examiner citations, which is more typical for validity searches, whilst face limits when analyzing a new published application, which includes only applicant citations, whose relevance can be limited,

and even more when starting a query from a patent draft, from a news or from a paper, which does not include citations.

Solution based on prototype text based queries (named "semantic" in patent informatics) are intended to overcome this limit, but this means also that the problem to solve is harder than citation-based solutions. In any case, semantic solutions proposed so far are so different together and moreover they are still evolving, that we have to analyze their features in more detail, to characterize the most typical implementations and also to identify technological trends, as we will do in the following paragraph. The recent reviews at recent patent events like CEPIUG 2018 [1] and EPOPIC 2018 [2, 3] identify a mixed feeling of today patent professionals on semantic searches, and some concerns, mostly due to the fact that semantic searches are perceived as black box solutions.
This contribution will try to "open the box" of semantic searches, summarizing first different typologies of semantic searches. Moreover, it will discuss why to use them and how to assess them, in order to facilitate the identification of user specific requirements.

1.2 Semantic search features and families

The table below summarizes the typical features of semantic patent searches, and the most typical alternatives for any feature.

n.	Feature	Alternatives	Kind of fature
1	Supported search	Semantic only search, semantic or other , semantic and other	User interface
2	Reference text	Full patent, patent passages, other text (e.g. a paper)	User interface
3	Input refinements	Not supported / possibility to identify or confirm key topics	User interface
4	Domain definition	Not supported / possibility to specify the technical domain	User interface
5	Results refinements	Not supported / possibility to reoder results / possibility to reactivate the query	User interface
6	vocabulary used	No vocabulary (LSA), Generalist or technology specific vocabularies	Internal implementation
7	semantic used	Words, concepts	Internal implementation
8	Supported language	English only, *Multilingual*	User interface, advanced
9	Kind of search	Prior art, *validity, freedom to operate*	User Interface, advanced
10	Text representation	Single bag of concepts, *Different concept bags, local text relationships, global text relationships*	Internal implement, advanced
11	Other info besides text	No other information, *metadata (IPCs/CPs, dates, source PO, target territory), citations*	Internal implement., advanced

Table 2: Semantic search features

Most of these alternatives are available in today implementations, others are cited as interesting alternatives, but not yet available: these are identified in Italic in table 2. For example, it should be interesting to have semantic multilingual searches, but today semantic searches are typically provided for English.

Features are further distinguished between those associated to the user interface, which are of course easier to identify directly, and those associated to the internal implementation, which can be eventually inferred from the data sheet or from information provided by the vendor.

The first feature to consider in assessing a product is the kind of search available (feature number 1 in table 2). Some products in fact provide a semantic search only, others add the semantic search to other more usual searches, in such a way that the user can select to kind of search to activate, e.g. to integrate results found in a usual search with those found in a semantic search. Another possibility is also to mix a more usual search, e.g. for identifying a first list of results, with a semantic one, e.g. to extend this list with other similar results, to increase the recall.

Other than this, products can be differentiated by features available in the HMI: these features are characterized by numbers 2 to 5 in table 2.

In some products it is fact possible to provide user selected passages, besides the full text (feature 2), or it is possible to identify or confirm key topics in these passages (feature 3). This allows the user to provide a more general input than a whole patent, and in any case to specify better his/her request, including the possibility of specifying the technical domain of interest (feature 4).

It has also to be mentioned the possibility of manually identifying the best results found in the first query and apply them for a second query (feature 5), a best practice in information retrieval used by the Rocchio algorithm [4, 5], which typically produce better results.

To summarize, from the user interface features mentioned so far, it seems that commercial solutions evolved from those available at the beginning of this decade, which attempted to provide almost automatic solutions, with a minimal human input, to those available now, which accept and even encourage the user to specific his/her request and feedback to obtain better results. Given that, I suggest to distinguish two generations of products, i.e. semantic searches 1.0 and semantic searches 2.0, characterized also by somewhat different objectives.

This distinction between semantic searches 1.0 and 2.0 applies also to internal implementation features, although these features can mainly inferred from the vendor literature.

First generation searches in fact emphasized the use of Latent Semantic Algorithms (LSA) [6,7] to identify related patents also if they do not use the same words of the reference text, but at least a set of them. LSA is a statistical method available since late '80s [8] which allows to identify related texts, without relying on the true meaning of concepts included in the text: this approach of course increases the recall, but results obtained this way are noisier than expected in patent searching.

The extraction of concepts, which can be expressed as a single word (e.g. "drone") or as a word sequence (e.g. "unmanned aerial vehicle"), and the inclusion of a suitable vocabulary which identifies that these expressions correspond to the same meaning should provide better results accuracy and is more common now in second generation semantic searches.

The table 2 mentions also other features, identified as "advanced", since today products typically provide the baseline alternative, but also other alternatives, identified in italic in table 2, are welcome.

For example, today solutions are available for English document only and are more suitable to prior art searches [1], whilst it is well perceived the need to use semantic searches for multilingual documents and also for other kind of searches.
These evolutions, when available, will produce a new generation 3.0, which will be enabled also by some advancement of the internal algorithm, as in features 10 and 11 of table 2.
To summarize, semantic patent searches 1.0, available since late '00s: 1) provided almost no possibility to detail the user intentions, and 2) relied on an algorithmic approach (LSA) to identify a shallow similarity in a whole text: this approach is adequate e.g. for searching news, but not satisfactory in patent searching.

Semantic patent searches 2.0 available today: 1) allow the user to identify better his/her intentions, and 2) relies also on vocabularies to identify the meanings of concepts in a patent text. Semantic searches 2.0 are more adequate for patent searching, although presently limited to prior art searching in English.

1.3 Using semantic searches

1.3.1 Why to use: to increase the recall

The risk of missing important results is always present in patent searching, and it is becoming more and more relevant with the continuous increase of the amount of patent applications by year.

This risk can be reduced by using Semantic searches, which should increase the recall.
In any case, semantic searches are typically less precise than "traditional" searches.

1.3.2 How to use: together with a "traditional" searcher / instead of a "traditional" searcher

Given that a semantic searcher has typically a higher recall and a lower precision than a "traditional" searcher, the preferred best practice is to use a semantic searcher to complement results of a traditional searcher, and to focus the analysis of semantic searcher results to those not yet found in the first step implemented by a traditional search.

More specifically, results of a semantic search after a traditional search can be categorized as in table 3.

Set	Definition	Note
A	Already found	
B	Not found before and important	This is the real added value of semantic searches
C	Not found before and useful	This category can also add some value
D	Not found before and to discard	

Table 3: Semantic search results categories

*Hence, semantic searches add a value when used together with a tradit*ional searcher, whilst do not provide *convincing* results *when used instead of "traditional" searchers, as reported also in recent evaluations [9], which in any case are focused to measure the precision, which is the typically weakest point of a semantic searcher.*

1.3.3 Which has to use a semantic searcher: skills required

The first generation of semantic searches provided almost no possibility to detail the user intentions, hence were also advertised as a way to automatize searches. In any case also in this case a professional user is useful to analyze the quality of results.

The second generation is more fair to the system and for the user, since it allows the user to specify his/her intentions, based on which the system can provide better results.

Hence, also "semantics searches" can't be used at best without the activity of a patent search professional.

1.4 Conclusions

The conclusions of this chapter hence are:

a) *Do not infer the characteristics of this technology by evaluating only some products: products available are in fact very different.*
b) *Try to assess products from data sheet first, by using the table 2 mentioned.*
c) *Semantic searches today justify their value in any case, since they increase the recall if used after a traditional search. This is the fairest way to evaluate them.*
d) *New generation and more accurate semantic searches require more user's interactions , hence their results depend on the user skills.*
e) *The technology is still evolving, hence be ready to a new wave of semantic searchers 1) multilingual 2) more accurate.*
f) *The search is only a possible of patent tasks which can benefit from semantics [10]: the next coming chapter 2 will present analyses oriented solutions relying on semantics.*

References

1) "CEPIUG 10th Year Anniversary Conference, Milano, Italy, September 2018" by Susanne Hantos, World Patent Information, Volume 56, March 2019, pp. 60-63 sections 2.3 "Practicalities of Semantic Searching" and 3.6 "Artificial Intelligence and the patent information business".

2) "Semantic search versus searching with terms and classifications. What helps best in patent information searching" presentation by Gabriele Kirch Verfuss, EPO Patent Information Conference, Brussels, 13 november 2018, accessed at https://www.epo.org/learning-events/events/conferences/2018/pi-conference/programme.html *on 24/3/2019*

3) "Black Box Patent Tools – Hope or Hype" presentation by Andrea Davis, EPO Patent Information Conference, Brussels, 13 november 2018, accessed at https://www.epo.org/learning-events/events/conferences/2018/pi-conference/programme.html *on 24/3/2019*

4) "Rocchio Algorithm" , from Wikipedia, accessed at https://en.wikipedia.org/wiki/Rocchio_algorithm *on 2/4/2019*

5) "Introduction to Information Retrieval" by Christpher Manning, Prabhakar Raghavan, Hinrich Schuetze", Cambridge University Press (2008), "Relevance feedback and pseudo relevance feedback", par. 9.1, pp. 163-172

6) "Latent Semantic Analysis" Wikipedia review, accessed at https://en.wikipedia.org/wiki/Latent_semantic_analysis on 02/04/2019)

7) "Foundations of statistical natural language processing" by Christopher Manning and Hinrich Schuetze, published by MIT (1999), par. 13.4.3 "Latent semantic indexing in IR", pp. 564-466

8) "Computer information retrieval using latent semantic structure" by Scott Derwester and alii, Bell Communication Research, applicant Patent US4839853A, US priority 1988, september 15, US grant 1989, june 13, retrieved from https://patents.google.com/patent/US4839853A/en

9) "AI in quest for an easier search", presentation by Burkhardt Schlechter, CEPIUG Conference, Milano, 9-11 september 2018

10) "Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics" by D. Bonino, A. Ciaramella, F. Corno in World Patent Information, n.32, March 2010, pp. 30-38