INNOVATION MINING
Supporting Web Mining in Early Innovation Phases

Jan Finzen and Maximilien Kintz
Fraunhofer IAO, Nobelstreet 12, 70569 Stuttgart, Germany

Keywords: Innovation management, Open innovation, Web searching, Semantic search.

Abstract: Fraunhofer IAO conducted a study among 1,000 German innovation professionals regarding their web-based information acquisition needs. The study showed the need for search tools and methods optimised for the target group of innovation professionals. In this paper we deduce accordant concepts from the study's results. We suggest an "Innovation Mining Process" as a structured approach to web-based information acquisition for early innovation phases. Our software prototype - the Innovation Mining Cockpit (IMC) - picks up essential concepts of this process and implements them as an easy-to-use web portal. The IMC is intended as a central point of contact for innovation-related search activities.

1 WEB SEARCHING IN INNOVATION MANAGEMENT

Innovative ideas can be the result of both formal and unstructured search processes and can have many different origins. The Internet provides access to external innovation sources in multiple and comfortable ways. However, the analysis process remains a complicated one: it is often unclear where the relevant information is located. Furthermore, classical search engines do not offer a sufficient precision to filter the results and separate the relevant ones from the large amount of irrelevant ones.

In a 2009 survey among innovation professionals, Fraunhofer IAO analyzed the Web searching requirements of innovation professionals in Germany (Finzen, Krepp and Heubach, 2009). Figure 1 summarizes the findings regarding the importance of different Web-based information sources regarding different steps of the early innovation phases: "innovation push", "idea collection", "idea creation", and "idea evaluation". While online journals and research and technology portals are used for finding innovation impulses and collecting ideas, encyclopedias and especially patent databases are most useful for idea evaluation. Internet-based information sources are considered most useful for collecting ideas, but less useful for the actual creation of ideas.

Internet-based information sources	Phases of innovation process				
	1	2	3	4	P_i
Search engines	36%	70%	35%	37%	16%
Online encyclopedias	13%	28%	16%	32%	8%
Online journals	32%	35%	15%	18%	9%
Research and technology portals	39%	32%	23%	18%	10%
Patent databases	22%	19%	14%	48%	9%
Economic databases	11%	16%	8%	28%	6%
Press release distribution services	28%	25%	10%	9%	7%
Expert finding platforms	8%	10%	9%	9%	3%
Open innovation portals	19%	12%	13%	8%	5%
Social networks	15%	13%	9%	12%	5%
Expert blogs	14%	14%	10%	18%	5%
Online market research	25%	31%	14%	27%	9%
Online forums	13%	13%	7%	13%	4%
Consumer blogs	11%	8%	6%	10%	4%
Percentage per innovation phase	26%	30%	17%	26%	

Top ten Others

Figure 1: Importance of Internet-based information sources for different (early) innovation phases (n=142).

We questioned the respondents about the most annoying problems they encounter with search engines. More interesting than the obvious aspects quality and time-efficiency however seem the remaining ones: almost half of the respondents claimed that they miss a ranking according to up-to-dateness and almost 40 percent are not satisfied with the available filter mechanisms.

Finzen J. and Kintz M..
INNOVATION MINING - Supporting Web Mining in Early Innovation Phases.
DOI: 10.5220/0003276202410247
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 241-247
ISBN: 978-989-8425-51-5

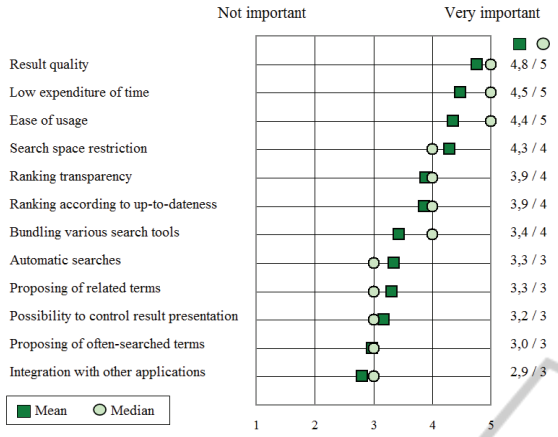

Figure 2: Importance of search engine features (n=142).

The results clearly show that there is a demand for a new generation of search tools tailored to the needs of innovation professionals.

2 INNOVATION MINING CONCEPTS

From both the results of our survey described above as well as from discussions with industrial customers we deduce several requirements that innovation mining tools must take into consideration. These requirements are discussed in the following sections.

2.1 Incorporate User Knowledge

Innovation professionals are considered to have a deep domain competence and bring along a broad knowledge of where and how to find information meeting their requirements and needs ("source competence"). It seems promising to make use of both the domain and the source competence as much as possible within a professional information acquisition tool, e.g. by

- allowing the users to integrate their favourite information source with little effort, and
- integrating domain knowledge into the search process. This can be done, for example, by hierarchically or ontologically structuring domain knowledge and providing both search queries and result sets for respective nodes.

2.2 Deal with Multitude of Relevant Information Sources

As shown in Figure 1 there are many kinds of information sources that are relevant to innovation management. Depending on the actual phase of the innovation process, some of these information sources are more important than others. According to our survey, the majority of respondents liked the idea of having multiple information sources integrated and accessible via a unified point of access. We divide the information sources available on the web into three different levels:

1. *The Document Level*: On the most basic level, single documents are accessed directly. E.g., the Fraunhofer website can be accessed by the user assuming to find the site at www.fraunhofer.de.

2. *The Database or Search Engine Level*: For most types of documents, specialised search tools already exist that offer advanced functionality to retrieve and present information. E.g., there are numerous search engines for scientific content, and web-based patent databases make it easy to search for patents on a given subject. To integrate such sources, a tool has to "speak" and "understand" the same "language", i.e., know how to formulate and send a search query to the search engine and how to grab and interpret the results. This is commonly known as "meta search".

3. *The Meta Search Level*: For some kind of information, meta search engines already exist. Accessing meta search engines requires basically the same requirements and process steps as for "normal" search engines, but is has to be kept in mind that one relies on the "language translation step" being executed by the meta search engine and thus not being under one's own control.

One challenge of providing an integrated information acquisition tool for innovation mining thus lies in the handling of "different languages" of different information sources.

2.3 Use Document-specific Metadata

There are many different document types relevant for innovation mining process, like e.g. patent data, scientific papers, press releases, or blog entries. Each of these document types has specific attributes and metadata that can be exploited within innovation mining: Press releases always have a publishing date and address information which make it easy to order and visualize them accordingly (Finzen, Kintz, Koch and Kett, 2009). Scientific literature can be a basis for expert identifiaction using co-authorship analyses. Patent data can be exploited for e.g. white spot analysis (Siwcyk, 2009).

2.4 Improve Ranking Mechanisms

One very clear result from our survey was the immense importance of ranking algorithms: People are quite unhappy with today's search engine's result ranking methods being only partly transparent and only poorly adaptable to the user. For innovation professionals, especially a ranking according to up-to-dateness and regional aspects showed up to be important. An information acquisition tool for innovation professionals thus should provide adequate possibilities. The determination of up-to-dateness, of course, is a challenging task and heavily depends on the document being analysed: Certain documents, like blog entries, patents, or press releases usually contain appropriate metadata. If explicitly marked, e.g. using an appropriate HTML or XML-tag, identifying them is trivial. Sometimes, however, they have to be extracted from the text – which is rather straight-forward by using regular expressions. In case of arbitrary web documents, the task is much more difficult: though timestamps may be found in the document it is not always obvious if it denotes the point in time the information in the text refers to. If no timestamp is given at all, the freshness might be estimated by comparing the content with another version of the content that has been indexed the last time the page was visited. Web monitoring tools do exactly that: they visit a web page in certain intervals and detect certain changes. By integrating carefully selected web monitoring patterns, a search engine might consider such documents that have been recently changed in the ranking method.

2.5 Support Long-term Information Needs

Search queries are often classified as being either

1. navigational (searching one or more specific document(s)),
2. informational (seeking information for a given topic), or
3. transactional (perform a particular action),

depending on the user's intention (Broder, 2002, Manning, Raghavan and Schütze 2007).

Almost 50 percent of all queries that general purpose search engine users utter belong to the first class (Broder, 2002). Consequently, current general purpose search engines mainly address navigational information needs. Nevertheless, the information needs of professional end-users (like market researchers, innovation professionals, etc.) often fall into the second or third categories. Springer (2006) stated that „the further development of tools to enable the detection of trends and the finding of information in the Internet can account for improving the innovation performance of companies".

Recurring searches are widely used within professional information management (cf. Finzen et al., 2009). The possibility to save and automatically repeat complex queries thus should be combined with effective ways to notify the user on newly found results. According to our study's results, techniques like RSS-feeds are still considered less important than more traditional communication channels like e-mail. Nevertheless we expect that the acceptance of such techniques will grow in the future, as they offer good means of integrating search results with further applications (e.g. knowledge management systems).

2.6 Offer Advanced Interaction and Visualization Concepts

Common use cases in innovation management include patent mining, competitor observation, and trend monitoring. To support these use cases information must be either extracted from one or more documents, or interpolated given an amount of documents. Such information needs do not only require special result presentation techniques, but also affect the query frontend: the users have to clarify that they do not want to be confronted with a list of documents but rather with, e.g., information extracted from documents ("new products of competitor X") or statistic information ("pie chart comparing the positive and negative utterances of forum users who wrote about company X in the last month", "bar chart showing the trend for a recent search topic for selected companies"). Search tools for professional end-users thus require suitable navigation concepts and powerful user interfaces for both, search query formulation and result interpretation.

2.7 Foster Integration and Collaboration

Web-based information gathering is a very individual task even in professional information work. However, with the size of an organisation the need to exchange search artifacts rises. It is quite common for larger companies to outsource search tasks to a special department (Finzen et al., 2009). With people searching for information together, the

need for methods and tools to foster collaboration arises. Ways to achive this include: exchange of bookmarks, persistable search spaces, sharing of search results and analysis reports.

3 INNOVATION MINING PROCESS

Building on the "tech mining" process described by Porter and Cunningham (2004) we suggest a five-step Web mining approach depicted in Figure 3.

1. *Identify Information Need:* Although the outcomes of any mining process might be surprising (after all, the idea of data mining is the detection of previously unknown facts), it should be as directed as possible. This means that the overall strategic goal of the mining process should be defined as the very first step, as it influences subsequent steps like source selection or visualization parameterization.

2. *Collect Information*: Depending on the information need, a variety of information sources is selected for mining. If the information need embraces temporal developments, e.g., trend analysis of a given topic, the document corpus must be created over a larger timeframe. However, for the case of innovation mining, it is usually important to gain information as soon as possible, i.e. ideally in real-time – as soon as they show up on the web. Therefore, either sophisticated crawl-and-scrape approaches or (even better) a push supply of information is needed.

3. *Process Results*: Once the source corpus is defined (and most probably being frequently expanded), the result processing starts. Though the approaches applied might vary very much, the main task of this step is a matching of a specified information need (e.g., a search query) against the documents of the corpus. In navigational search the tasks usually ends with weighting the respective document's relevance in relation to the information need. This allows a suitable ranking of documents in a subsequent step. Information needs that are of a more informational kind typically involve additional information processing steps: information or metadata extraction forms a basis for appropriate analyses in the subsequent step.

4. *Analyze and Interpret*: The information collected in the processing step are analyzed and condensed, and finally put into graphs that suit the information need. This step aims at supporting the analysis of the data by the user as good as possible. It includes choosing the best-fitting visualization, the tailoring of the visualization regarding the results to display as well as possibly providing additional information that helps interpreting the data in the right way.

5. *Disseminate and Act*: Once interesting results are found, subsequent processes can be triggered: further results showing up in the future might be automatically taken into account and thus resulting in new versions of the result analysis. When new findings are available, the user might want to be notified by an appropriate alerting mechanism. Results might be saved and reloaded, printed, exported into complementary software tools, and passed on or shared to other users

4 INNOVATION MINING COCKPIT

To evaluate our deductions regarding search tool requirements of innovation professionals, we implemented a search engine prototype. The Innovation Mining Cockpit (IMC) picks up essential concepts of this process and implements them as an easy-to-use web portal.

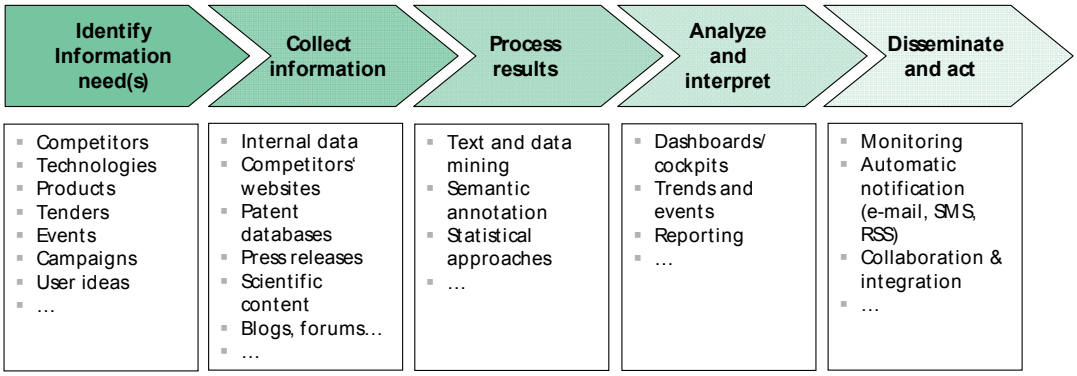

Figure 3: Innovation Mining Process.

We will present the IMC's main features regarding the requirements discussed above in the following section.

4.1 Source Identification

The source identification module implements a meta search engine approach: The keywords entered in the text field are passed to various general purpose search engines. The results of the different search engines are combined, ranked, and displayed in a typical search results list. Either the complete URL or the Web domain can be added to the search space (the websites initially show up in a special folder "New sites" within the Search Space Configuration module). The Source Identification module aims mainly at step 2 ("collect information") of the Innovation Mining Process and is particularly important to quickly add new information sources to the mining process. In combination with the Search space configuration module described in the next section, the source identification module illustrates one basic design paradigm of the IMC. The search and mining process is restricted to such sources that the user considers potentially relevant. This approach, of course, not only lowers the number of irrelevant results in the search process but also restricts the resource requirements (regarding computation and storage hardware) significantly as opposed to a broad crawling approach. On the other hand, it makes finding and selecting the significant sources a most important preprocessing step.

4.2 Search Space Configuration

The Search Space Configuration module supports step 1 ("identify information need") and 2 ("collect information") of the Innovation Mining Process. It combines a sophisticated bookmarking system with several search engine specific adjustments, like crawl depth, support of named entity recognition, or automatic recognition of RSS feeds. Figure 4 shows a screenshot of the Search Space Configuration module.

It allows distinguishing between "normal" Websites and feeds that offer an information push supply. The main difference lies in the information supply paradigm. As RSS-based information offer several advantages compared to a classical website, we built the IMC's whole data aggregation mechanism on the feed principle. Based on some heuristics (like text block length), significant changes and new content are detected and provided to the IMC in an RSS-like format. Thus, the IMC internally builds its own feeds for any website that may not offer one explicitly. This way, the user can easily be notified on any (relevant) newly found information on any site either by RSS, E-Mail or SMS.

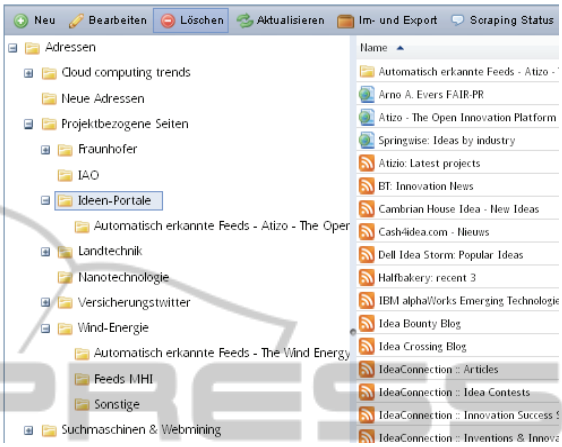

Figure 4: Search space configuration.

4.3 Feed Search

The feed search portlet offers advanced functionality to search and analyze feeds:

- Any configured search can be turned into an RSS or Atom feed using the respective buttons at the bottom of the dialog.
- The timeframe can be limited using the calendar pickers at the top of the dialog.
- A feed search configuration can be saved and loaded. This is especially important as configurations can become quite complex – consisting of source restriction, filter settings and the actual search query. This supports the long-term information needs of innovation professionals.

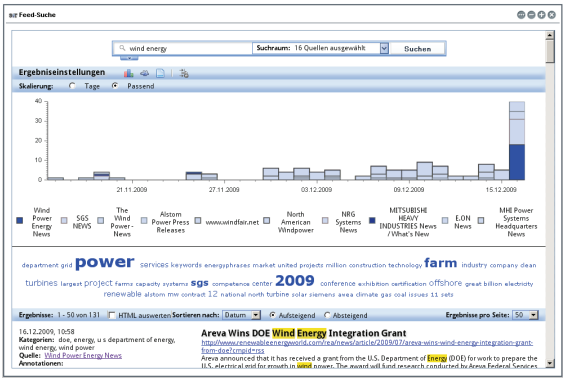

Figure 5: Feed search frontend.

Figure 5 depicts the feed search module showing the results for a "wind energy" query. The result list is accompanied by trend chart and tag cloud visualizations are provided to (i) provide additional information about "hot topics" and (ii) further navigate through the search result set.

4.4 Change Monitoring and Notifications

Information needs in innovation management are often rather informational than navigational and rather long-term than ad-hoc (Finzen et al., 2009). The Feed Search portlet therefore provides means to automatically execute the query in the background and notify the user on newly found results either by e-mail or using a feed reader. The e-mail notification report can be scheduled for any search either on a regular basis (e.g., once per day or week) or as soon as new results have been found (which of course depends on the interval the website's content is being compared by the IMC's crawler, or the feed polling interval configured in the Search Space Configuration module. E-mail adresses are configured within the portal server's user management system.

4.5 Semantic Annotations

Depending on the document type and the information source, a document may embrace metadataor even semantic markups which can be utilised to offer advanced result visualisations and browsing functionality, e.g. facetted search. Unfortunately the amount of semantically annotated web content today is still very limited.

The IMC integrates the OpenCalais web service (http://www.opencalais.com) on demand during the search space configuration process. As OpenCalais is currently not available in German, we also integrated the AlchemyAPI web service (http://www.alchemyapi.com) by orchest8, which offers similar functionality.

Figure 6 shows results in the Feed Search portlet that have been annotated via the OpenCalais web service. Countries and organisations are provided in the left meta data columns. Clicking on a meta data link restricts the current search accordingly, e.g. when clicking on "U.S. Department of Energy", only results that include a reference to this organisation will be displayed.

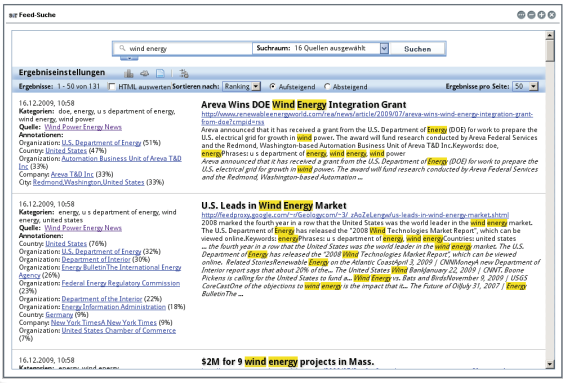

Figure 6: Semantic annotations.

4.6 Visualization

Special emphasis has been put on search result visualization.

Frequency Analysis: trend monitoring and event detection are important use cases within the area of technology and innovation management. The innovation mining cockpit uses tag clouds, classical bar and line diagrams for any possible search query.

Geographical Analysis: oftentimes geographical information can be extracted from texts fairly easily (e.g., country names can be easily maintained in look-up lists). For geographical information, the longitude and latitude can easily be assigned using respective geo-coding web services. This allows the visualization of objects on maps.

Association Analysis: association graphs are used to show and analyze relations between objects, e.g. companies or persons. Generally, the nodes represent objects and the edges the relations between these objects. Various layout algorithms help to analyze structural information within the graph. For example, the degree of cross linking may indicate a company's importance (or at least its activeness) within a technological area.

5 CONCLUSIONS AND OUTLOOK

The current version of the IMC has been presented to several industrial partners and obtained very encouraging reactions. We are currently running through a long-term evaluation with an industrial partner of the automotive sector to evaluate how the different modules are accepted in an innovation professional's daily work. Even though final evaluation results are not available yet, we already

achieved valuable feedback that will be addressed during short-term development of the tool:

- The current implementation forces users to index a whole website (using the main URL combined with a high crawl depth). This leads to a large amount of irrelevant content, as the main intention of the IMC is to mine only such content that is very fresh. In the next version of the IMC, a wizard will help users to identify the relevant parts of a website by generating a sitemap of a website and let the user select branches of this maps, e.g., company news, press releases, discussion forums, etc.
- Our association analysis is currently solely based on co-occurrence analysis. This proves useful in selected use cases such as "who knows whom" or "who works with whom" analysis but demands for a very well selected source corpus, as results can easily be polluted by a misplaced source containing lots of named entities, like e.g., a stock report mentioning 100 companies. Although using thresholds regarding the occurrence number proves very helpful to get rid of such problems, combining the co-occurrence approach with more sophisticated approaches based on linguistic and semantic analysis seems promising.

Table 1 summarizes further needs and plans for future extensions.

Table 1: Future work.

Requirement	Future work
User knowledge	Semantic domain models to utilize user domain knowledge
Integration	Specific support for scientific literature, patent data, forums & blogs, and open innovation portals
semantic annotation	Use uniform annotation framework based on Apache UIMA (http://uima.apache.org)
Ranking	Improve configurability of relevance parameters
Collaboration	Reports, private and public spaces

More information on Innovation Mining can be found on http://www.innovation-mining.net.

ACKNOWLEDGEMENTS

The project was funded by means of the German Federal Ministry of Economy and Technology under the promotional reference "01MQ07017". The authors take the responsibility for the contents.

REFERENCES

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum* 36, 2 (2002): 3-10.

Finzen, J., Kintz, M., Koch, S., Kett, H. (2009). Strategic Innovation Management on the Basis of Searching and Mining Press Releases. P*roceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST)*. Lisbon.

Finzen, J, Krepp T, Heubach D. (2009). Web Searching in Early Innovation Phases: a Survey among German Companies. *Proceedings of the 2nd ISPIM Innovation Symposium.* New York.

Manning, C. D., Raghavan, P., and Schütze, H. (2007). *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England.

Porter, A. L., and Cunningham, S. W. (2004). Tech Mining: Exploiting New Technologies for Competitive Advantage. *Wiley Series in Systems Engineering and Management*. John Wiley & Sons.

Springer, S. (2006). Nutzung von Internet und Intranet für die Entwicklung neuer Produkte und Dienstleistungen" *nova-net Werkstattreihe*, Stuttgart: Fraunhofer Verlag.

Siwczyk, Y. (2010). *IT-gestützte White-Spot-Analyse: Potenziale von Patentinformationen am Beispiel Elektromobilität erkennen*. Stuttgart: Fraunhofer Verlag.