SustaInno: Toward a Searchable Repository of Sustainability Innovations*

Walid Shalaby, Wlodek Zadrozny
Computer Science Department
University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223

Abstract
In this paper we describe our ongoing work on SustaInno; an open-source search repository of innovations related to sustainability. SustaInno utilizes advanced information retrieval and text processing methods on technical innovations (initially patent data) to provide its users with practical, applicable, and detailed solutions to their sustainability related challenges. For example, problems like urban heat islands and rainwater waste are of major concern to most urban cities. Using our repository, decision makers can get quite in-depth solutions on practical approaches to address these and many other problems. The novelty of our work stems from three main factors: (1) such a repository does not exist, (2) it is focused on sustainability innovations which are of great importance for the creation of sustainable living environment, and (3) it provides a set of open-source tools and open-access datasets that could accelerate the dissemination of knowledge about sustainability.

Introduction
With the continuous increase in human population, our environmental needs grow very rapidly. Given the fact that we have limited resources on our planet, future generations might face lack of resources. Nevertheless, with sustainable development we could alleviate that problem to large extent. Having said that, we believe the society awareness of sustainability is still bounded. Additionally, there is a lack of utilization to existing technical knowledge sources that address sustainability related challenges (e.g., technical innovations).

Different initiatives were proposed to popularize society awareness and regulate decision makers’ actions toward environmental, economic, and social sustainability issues. For example, the US Environmental Protection Agency (EPA) (EPA 2014) was instituted specifically to prepare and enforce a set of regulations in order to protect the environment as well as public health. Nevertheless, with deeper inspection of these regulations, one can realize they are general guidelines without any technical solutions to the given problem.

For instance, consider the problem of Urban Heat Islands (UHI). EPA through (EPA 2013) provides mitigation strategies like trees and vegetation, green roofs, cool roofs, and others to address this problem. These strategies are more detailed (in their corresponding webpages) by giving their impact as well as benefits and costs. Moreover, (EPA 2013) provides links to federal and state actions webpages which further give details about initiatives as well as plans and design guidelines to address the impacts of UHI.

As can be noticed, though these data are of great importance and value from general knowledge perspective, they represent general guidelines lacking specific solutions with information on how to apply them in reality. These restrictions and others motivated us to leverage other unexplored technical knowledge sources like patents and examine their applicability to the sustainability domain.

This paper describes our ongoing work on SustaInno; a search repository of technical innovations related to sustainability. We chose technical innovations (patents) as the underlying data source because they provide practical, applicable, and detailed solutions that consider environmental and economic factors, hence of sustainability flavor. Furthermore, technical innovations are of evolving nature (WIPO 2014) meaning that they will always be tailored to contemporary and recent challenges. More specifically, we used a subset of the US patents and patent applications between 1976 and 2013 derived from a list of topics related to sustainability.

Work with patent data as source of innovative solutions to sustainability issues was not reported before. In the meantime, patent information retrieval was addressed in detail in (Tait and Trippe 2011). Most research in patent retrieval was general purpose meaning it didn’t address special characteristics of specific domain or technology. Some research efforts were proposed targeting specific patents category like chemistry patents (Lupu, Huang, and Zhu 2011), or chemical-structure searching in patents (Holliday and Willett 2011). Our work represents an extension to task-based patent retrieval by creating a searchable repository of sustainability innovations. With this in mind, we tailored our repository to be problem/solutions oriented so that users can express their search queries as problems and obtain corresponding solutions latent in innovations text. This required adding semantic elements to our repository that identify tex-
tual parts describing solutions to a particular problem.

We also tailored the repository as a search engine using open-source tools and libraries to serve our objective of allowing different stakeholders like decision makers, policy planners, researchers, and even the public to easily access and interact with our system. This will help alleviating two main problems: 1) society need for solutions addressing sustainability issues e.g., energy consumption/generation, and 2) data science need for exploring, deeper understanding, and providing open-access tools of technical innovations.

**Sustainability Innovations Collection**

As a first step toward the searchable repository creation, we firstly sought collecting innovations related sustainability topics. Our main data source of innovations was the US patent office data hosted by Google (Google 2012). We started by crawling all patent grants and patent application publications between 1976 and 2013. We extracted all relevant fields including: title, abstract, description, claims, US classification, publication number, publication type, publication date, filing date, and assignee.

To construct the sustainability innovations pool (i.e. the set of patents that are related to sustainability), we followed a naive two-step process: (1) we stored all relevant topics in Wikipedia index of sustainability articles (Wikipedia 2014) in a lookup table resulting in a total of 100 topics, and (2) from the full collection of crawled patents, we added a patent to the pool only if its title contains one of the entries of the lookup table created at step 1. This process resulted in approximately 33,400 unique patents divided among the topics, where some of these patents have mixed membership to more than one topic. Among the top 20 topics in terms of number of innovations in each topic there are solar technologies, hydrogen technologies, recycling, photovoltaic, wastewater, and others (see table 1). Frequencies of innovations of the top 10 sustainability topics are given in figure 1.

<table>
<thead>
<tr>
<th>solar</th>
<th>electric vehicle</th>
<th>geothermal</th>
</tr>
</thead>
<tbody>
<tr>
<td>hydrogen</td>
<td>hybrid vehicle</td>
<td>solar heating</td>
</tr>
<tr>
<td>recycling</td>
<td>emission control</td>
<td>energy efficient</td>
</tr>
<tr>
<td>solar cell</td>
<td>heat recovery</td>
<td>energy saving</td>
</tr>
<tr>
<td>photovoltaic</td>
<td>water purification</td>
<td>low energy</td>
</tr>
<tr>
<td>wastewater</td>
<td>biomass</td>
<td>solar power</td>
</tr>
<tr>
<td>wind turbine</td>
<td>wastewater treatment</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Top 20 sustainability related topics in terms of number of innovations (ordered top-down then left-right)

It is worth mentioning that, this simple process could be extended in many ways to allow more patents to be added to the sustainability innovations pool during the filtration process. For example, we could utilize patent fields other than the title (e.g. abstract) as a reference to the lookup operation. We could also expand each topic patents by following a majority voting scheme using their classification codes; thus, if the majority of patents under specific topic belong to specific classification code(s), we might add all other patents having same code(s) to the sustainability innovations pool.

Moreover, we could utilize more advanced information retrieval techniques for sustainability innovations expansion. For example, to expand the list of sustainability topics, we could train topic modeling techniques like LDA (Blei, Ng, and Jordan 2003) on a set of sustainability related articles and resources. We could also train a classifier on the current set of patents in the sustainability pool and use it to predict whether other patents are sustainability related or not.

The Search Repository

Our search repository is composed of multiple modules and sub-modules. Figure 2 shows the high level architecture of our repository where two main modules are continually running: the index generation and search. Below we give detailed description of how each of these modules is working and integrated into the search repository.

**Innovations Index Generation**

We used Apache Lucene™ (Apache 2011a), an open-source java library, to create our repository index. As shown in figure 2, the index generator uses the collection of patents in the sustainability innovations pool to generate and organize the index such that, patents are mapped into documents and patent fields extracted during data crawling are mapped into document fields. This fine grained representation of index elements allows users to search the index using either general text queries or queries about specific patent field(s).

As we plan to develop a semantic search repository, we initially added to the index some semantic fields indicating problems addressed by the innovations and their corresponding solutions. The purpose of these fields is to allow the user to express his search query as a problem he is facing and in return display search hits as solution(s) to the given problem as expressed by the innovation. For generating these fields, we used simple regular expressions that match sentences containing patterns like "present invention", "advantage", "disadvantage", "problem", "mitigate", and others.

It is worth mentioning that, this shallow semantic processing of patent text allowed us to identify the contribution of a particular innovation and contrast it with the alternatives discussed in the prior art. As we continue developing the repository, more semantic elements will be added to the index through deeper text understanding.

**Search Integration**

As we aim to provide a set of open-source tools for technical data exploration, we used Apache Solr™ (Apache 2011b) as the underlying search engine of our repository. Solr is an open-source enterprise search engine based on the Lucene project.

Solr is highly customizable through its configuration files which allowed integrating our index very easily. As shown in figure 2, we leveraged additional Solr features like faceted search and clustering to give the user a more organized view of the search results hence enhancing his experience with the system. We also configured the search hits scoring function...
Figure 1: Innovation frequencies of the top 10 sustainability related topics

Figure 2: The search repository architecture
such that the index semantic fields (i.e., problems and solutions) have higher match scores than normal text fields (e.g., title, abstract). This will guarantee popping up innovations with solutions to user queried problem in the search results.

Discussion and Future Work

This paper introduces our ongoing work toward creating a searchable repository of sustainability innovations. We chose sustainability as the initial theme of the repository for three reasons: (1) its importance to the society which would translate the success of the project into tangible benefits for the society, (2) its clear interest among different stakeholders which would give the project frequent feedback and eventually contribute toward making it more efficient and effective, and (3) such a repository will accelerate dissemination of knowledge about sustainability hence help creating better sustainable environment.

We built the repository using a subset of the US patent data which encode innovative solutions to many sustainability related problems. Our initial repository holds around 33,400 innovations addressing different topics like solar cells, recycling, hydrogen technologies, photovoltaic, wastewater, and others. This initial repository was built using a simple lookup operation of sustainability related topics in patents title.

There are many ways to expand this repository hence increasing its coverage. For example, by expanding the topics list using topic modeling techniques like LDA or simply going beyond (Wikipedia 2014) and inspecting other online sources. We could also use a more tolerant lookup approach which utilizes other patent fields like the abstract and description. Additionally, a classification model could be built from the initial repository and used to predict future patents related to sustainability.

Our search repository addresses the need for solutions supporting sustainability by utilizing more than just patents text. By incorporating semantic elements like problems and corresponding solutions into our index, we increase the system ability to respond to problem-oriented queries. It is true that we used unsophisticated regular expressions to extract problems and solutions elements, yet it is just a start toward fully semantic search repository by deeper understanding of innovations text. Our approach could be extended by parsing documents and dividing them into chunks, then applying shallow natural language understanding techniques to identify other semantic parts describing solutions to problems.

Another objective of this work was to provide a set of open-source tools and open-access datasets for exploring creativity latent in one of the largest text corpora; patent data. Through our repository, we realized this objective by utilizing open-source libraries like Lucene and Solr. We also plan to provide free online access to our repository by hosting it on a public domain. This will allow other researchers to test their methods on a large, cured, and dynamic textual dataset. With additional material, the open-source tools and datasets used in our work can be used in data science education.

References