Pancreatic Cancer Search Facilitator

David Ochoa[1], Iana Zankina[1]

[1]Graduate Students, Computer Science and Engineering, FAU

Abstract

Newly diagnosed Pancreatic Cancer Patients are naturally curious about their disease. Information can help them deal with the trauma of their diagnosis and help them make informed decisions about their course of treatments and options. Unfortunately, the web can be a vast and overwhelming challenge, especially for someone who is already feeling overwhelmed and alone due to a harsh and sobering cancer diagnosis. The goal of our Pancreatic Cancer Search Facilitator is to alleviate a lot of the stress related to inaccurate and spam like search results stemming from typical Internet search engines. Patients should get fewer and more accurate results to their searches with this specialized search engine made specifically for their topic of interest.

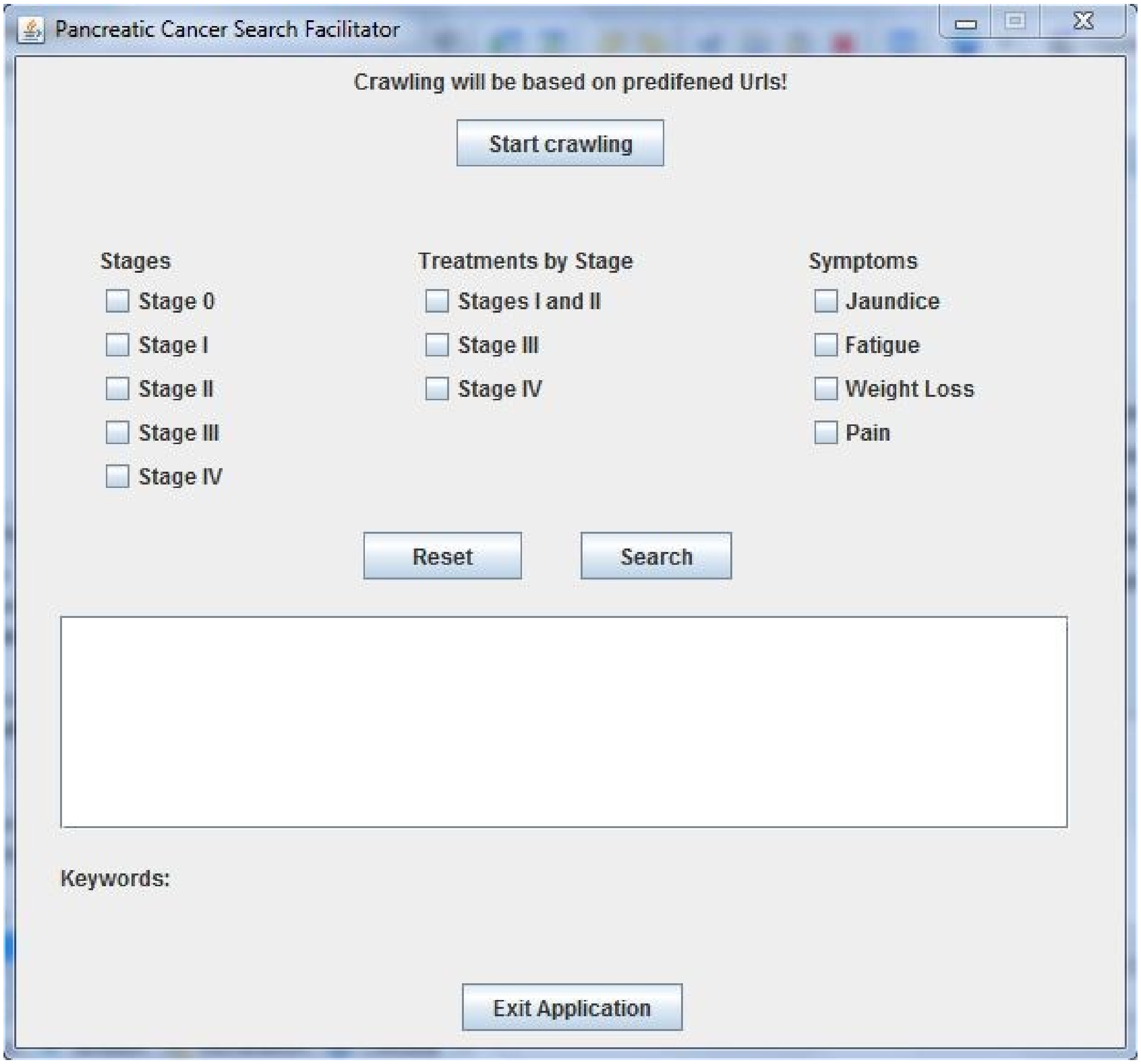
Keywords: Pancreatic Cancer, Stages, Treatments, Symptoms, Risk Factors, Prevention

Background

As we know and although there has been a lot of progress in the search for a cure for pancreatic cancer, the results are still not what we expect. In the same way as research is done on a daily basis, information keeps growing as a consequence. For that reason, we created an easy-to-use tool to help patients get accurate information when needed depending on their own particular case or for a love one. In the following paper we describe a user interface application for pancreatic cancer that helps patients diagnosed with such disease in the retrieval of specific data based on keywords such as stages, treatments, symptoms, etc.

Methods

The Pancreatic Cancer Search Facilitator (PCSF) is an easy to use applications created for the purpose of helping users get the right information when searching for pancreatic cancer as the main subject. In order to come up with the information, we have created an ontology, a graphical user interface, and web crawler and indexer capabilities (with help from code provided by our textbook [2] to accomplish our goal).



Crawling will be based on predifened Urls!

Stages
□ Stage 0
□ Stage I
□ Stage II
□ Stage III
□ Stage IV

Treatments by Stage
□ Stages I and II
□ Stage III
□ Stage IV

Symptoms
□ Jaundice
□ Fatigue
□ Weight Loss
□ Pain

Keywords:

Figure 1: User Interface

Some of the tools used in this project are Protégé and Eclipse (integrated with the Jena java framework for building semantic web applications). We chose Protégé to create our ontology because it is Java based, open source, and widely used by developers; on the other hand, Eclipse is a multi-language software development environment that includes an integrated development environment. An ontology is necessary as a way to define a common vocabulary for anyone who needs to share information in any particular way.

For example, in our case we have many different web sites containing medical information with respect to pancreatic cancer. If these web sites share and display the same ontology of terms they all use, the computer can add and remove information from all these different sites. In that way, we can use the information to answer user queries and come up with a better search result. Our ontology is essential to our application because it contains the different keywords that are used to narrow down which sites contain the pertinent information the user is interested in. For that reason it must be accessible in our Java application. This is accomplished through the Jena java framework code as well as query methods using SPARQL querying language. Integration of these different components is crucial for a powerful and necessary effect. In essence we are able to tie the ontology created using separate software, Protégé, to our Eclipse application.

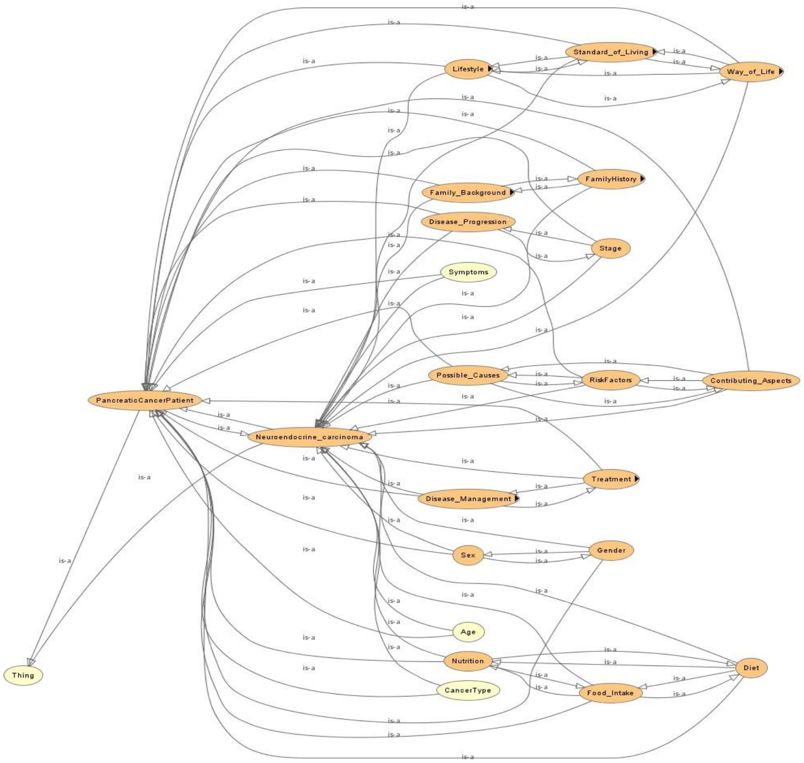


Figure 2: Representation of inferred Ontology

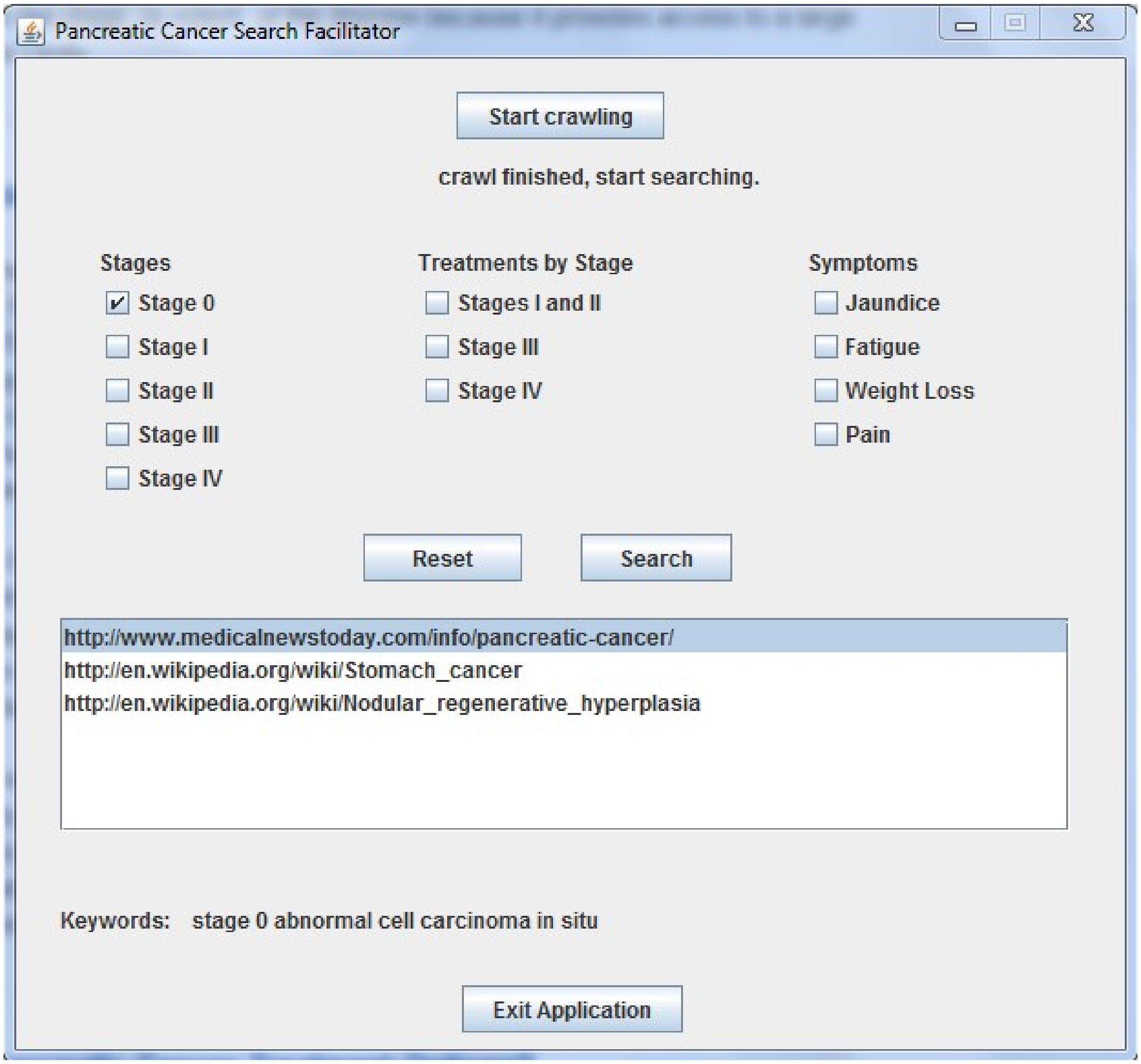
As mentioned before, we use Eclipse to create our graphical user interface which consists mainly of buttons, labels, checkboxes, and a list. Furthermore, Eclipse and the Jena framework, are used to integrate the ontology we have created with Protégé as well as to integrate the code given in the textbook from chapter 2 which is written in java. From all the java files contained in the book, Lucene in Action, there are three main files essentially important for our project; they are FetchAndProcessCrawler.java, LuceneIndexer.java, and MySearcher.java; we could have modified all of them but in our case we only modified FetchAndCrawler.java in some extent in order to achieve our goal. We use the FetchAndCrawler.java file in order to follow (crawl) hyperlinks throughout a Web site to retrieve the data and parse it in a methodical and automated manner, based on the URL's that we feed our program. Then we use the LuceneIndexer.java file to index our processed content, which has been parsed into text files, so the searcher can find it later based on various search criteria. Finally, we use the MySearcher.java file to search through our newly created index so we can find the information relevant to the search criteria based on keywords. Other files found on the packages iweb2.ch2.webcrawler.parser.html, iweb2.ch2.webcrawler.parser.common, and iweb2.ch2.webcrawler.parser.parser are used in the background to parse the information found on the web into text-based files for searching purposes. As explained before we made little changes to the FetchAndCrawler.java file as shown below. We basically remove some of the code we were not going to use and added our own URL's.

```
public void setAllUrls()
{
   setDefaultUrls();
}
```

```
public void setDefaultUrls()
{
  addUrl("http://www.cancer.gov/cancertopics/pdq/treatment/pancreat
        ic/Patient/");
  addUrl("http://en.wikipedia.org/wiki/Pancreatic_cancer");
  addUrl("http://www.mayoclinic.com/health/pancreatic-
        cancer/DS00357");
  addUrl("http://www.medicinenet.com/pancreatic_cancer/");
  addUrl("http://www.medicalnewstoday.com/info/pancreatic-cancer/");
  addUrl("http://www.webmd.com/cancer/pancreatic-cancer/");
  addUrl("http://www.nlm.nih.gov/medlineplus/pancreaticcancer.html");
  addUrl("http://www.pancan.org/");
  addUrl("http://pathology.jhu.edu/pc/");
  addUrl("http://www.lustgarten.org/");
  addUrl("http://www.pancreatica.org/");
}
```

Results

The application is easy to use; the first step in getting the results we want is to click on the button labeled start crawling. The purpose of doing so is to crawl as many links as we can, based on the URL's that we seed the program with, in our case pancreatic cancer information web sites. At this point the user does not know how many URL's have been found, they just have to wait until the crawling is done to continue to the next step. Second, and with the help of the checkboxes, the user selects if he or she is interesting in finding information regarding the stage, treatment by stage, or symptoms of pancreatic cancer. After the selection(s) has been made we click on search so we can get and display the links containing the information requested. The resulting links are populated based only on the user's options and to make notice of that, we created a label which will display specific keywords contained in the documents related to the user's selection. Finally, these links eventually redirect the user to such documents for further reading at the user's behest.



crawl finished, start searching.

Stages
☑ Stage 0
□ Stage I
□ Stage II
□ Stage III
□ Stage IV

Treatments by Stage
□ Stages I and II
□ Stage III
□ Stage IV

Symptoms
□ Jaundice
□ Fatigue
□ Weight Loss
□ Pain

```
http://www.medicalnewstoday.com/info/pancreatic-cancer/
http://en.wikipedia.org/wiki/Stomach_cancer
http://en.wikipedia.org/wiki/Nodular_regenerative_hyperplasia
```

Keywords: stage 0 abnormal cell carcinoma in situ

Figure 3: GUI results after crawling and indexing

Discussion

As of today the Pancreatic Cancer Search Facilitator works as we expected; it is easy to use and we are getting the result we want, and indeed the result that a future user will want while searching for information with respect to pancreatic cancer when the survival of a patient is at risk. In any case the application has much to offer as long as we keep improving it. Some of the improvements that we could add to the application to make it more efficient will be adding a text field. This would improve the application by allowing the user to enter the Web site he or she wants to crawl and gather information on, rather than the Web sites we are suggesting. Another addition will be adding a package to the main project that will contain the files necessary to parse pdf files. So far we do not use this capability it in our current project. Much of the information that has been crawled is in the html format, but a pdf parser will be a nice addition to the project. Another future improvement would be expanding the ontology to include more synonyms and classes of keywords. There is practically an endless amount of terminology related to Pancreatic cancer, and so there is a lot of room for growth in the associated ontology. Also this would mean our GUI should be extended to accommodate this growing list of options for the user to choose from. As with any project, after a certain level of capability is reached and the developer is satisfied with its completion, there is only one thing left to do, improve it!!

Conclusion

Eliminating the stress in web searching is beneficial for users. In the case of seriously ill pancreatic cancer patients, it is even more necessary. The goal of our project was to come up with an easy-to-use tool that populates a knowledge representation of accurate information on pancreatic cancer. The information is specific enough so as to allow users to filter out irrelevant data. Through the integration of different software and coding strategies, this application creates a purpose driven search engine, which gives the impression that it is capable of thinking as to what results the user expects and then acts accordingly to generate it.

References

[1] Hebeler, J., Fisher, M., et al., Semantic Web Programming, Wiley, 2009

[2] Marmanis, H., and Babenko, D., Algorithms of the Intelligent Web, Manning, 2009

[3] McCandless, M., et al., Lucene in Action, Manning, 2010

[4] Bellifemine, F., et al., Developing multi-agent systems with JADE, Wiley, 2007

[5] http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

[6] http://www.javaranch.com/journal/2004/04/Lucene.html