Project Idea Document

IN4325: INFORMATION RETRIEVAL

Group 13:

Jens Voortman	4081005
Mingcheng Song	4583396
Shashank Rao	4624017
Siyu Chen	4587456
Wouter van der Zwan	4019806

March 10, 2017

1 Introduction

The Problem

Nowadays, there are various lifestyle applications available in the market, offering convenient services from booking hotels, flight and train tickets to restaurant recommendation and reviews, like Booking.com[1], Tripadvisor[2] and Yelp[3]. People are willing to use these apps to help them look for and compare different items until they find satisfying results. However, no such website for health care exists in most countries, while health is one of the most important themes of human lives. Based on the above facts, it is a good idea to develop a medical search system providing people with more convenience.

Proposed Solution

The proposed solution is a website with a familiar search, compare and rank interface comparable to the sites mentions above. On this site users can type the name of a medical issue and get a localized recommendation on where to find the best medical care for that issue. The resulting hospitals will be ranked according to 'medical expertise' on the search illness. The 'medical expertise' of a hospital will be calculated from the number of clinical trials on the subject the hospital participates in and the number of publications by the hospital's doctors on the subject.

Challenges

Of course, this isn't easy to do, otherwise there would have been a company very rich by now. The first thing is to get a list of all hospitals with their official websites. Even though some sources like Foursquare[4] offer data in a structured manner, the data might be outdated or incomplete. Matching the correct URL to the correct hospital is not an easy task.

Then there is the search for doctors. All hospital websites are different, making the data completely unstructured. Some hospitals might have a list of their doctors hidden somewhere deep inside their website. And some will even not have such a page at all. Others might have their doctors hidden behind a search engine. The key challenges here are finding the correct pages and extracting the doctors names from the found pages.

Finally there is the extracting of the doctor's specialization. While ClinicalTrials is very structured with its data, PubMed is a lot less structured. Finding the medical keywords/keyphrases and mapping them to the doctor's specialization from a PubMed page might prove quite difficult. It would possibly require applying keyword extraction techniques on the publications. Existing tools for keywords extraction are: *RAKE* (statistical method Rose et al., 2010), *Maui & KEA* (supervised learning method, Witten et al., 1999) and *Word2Vec* (unsupervised learning technique, Li, Zhu, and Lu, 2015).

[1] https://booking.com

[2] https://www.tripadvisor.com

[3] https://www.yelp.com

[4] https://foursquare.com/explore?mode=url&q=hospital

2 Requirements

To assign priorities to various parts of the system we are going to develop, the MoSCoW model will be used. The MoSCoW model works with four priorities: Must have (essential for the system), Should have (important yet not essential), Could have (to be implemented if time permits) and Would have (might be interesting to look at in the future)

Must have...

- ... a method to find all hospitals in a location.
- ... a method to find a URL for a hospital if not provided.
- ... a method to find doctors from a hospital website.
- ... a method to find if a doctor has medical publications.
- ... a method to extract keywords from medical publications.
- ... a method to find if a doctor is involved in Clinical Trials.
- ... a basic interface for the user to query our system.
- ... a method to rank the hospitals based on their quality.
- ... respect for Web server policies.

Should have...

- ... a responsive interface for the user to query our system.
- ... a mapping capability of illnesses to specializations.

Could have...

- ... a Distributed Crawler Pipeline.
- ... an interface with a map with results.
- ... the ability to index all english-speaking hospitals.
- ... a search for doctors instead of hospitals.

Would have...

- ... support for languages other than english.

3 Pipeline

We can define the retrieval process in a couple of steps. First we need to gather the local hospitals from Foursquare. Once we have the list of hospitals we need to process them to see if they all have a URL to their website. If this is not the case, then we need to crawl the Web to see if we can find it. Once we have the websites, we can crawl them for doctors. Once we have a list of doctors that belong to each hospitals, we can crawl PubMed and ClinicalTrials to find the publications and trials information for each doctor. A visualized version of this pipeline flow can be found in figure 1.

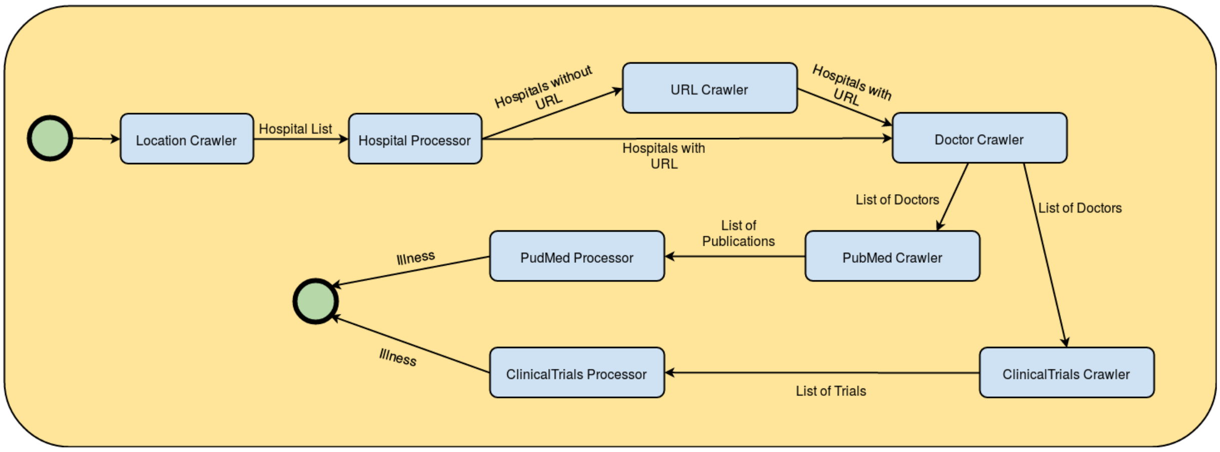


Figure 1: Abstracted Retrieval Pipeline

4 User Interface

Once we have the information, we need to make it accessible to the user in a UI. We have two mockups for them. The first one requires few time to create, but is very basic in its use. The second one is more sophisticated and has more features like the results shown as a map. However, as this one is more complicated, it will take longer to create.

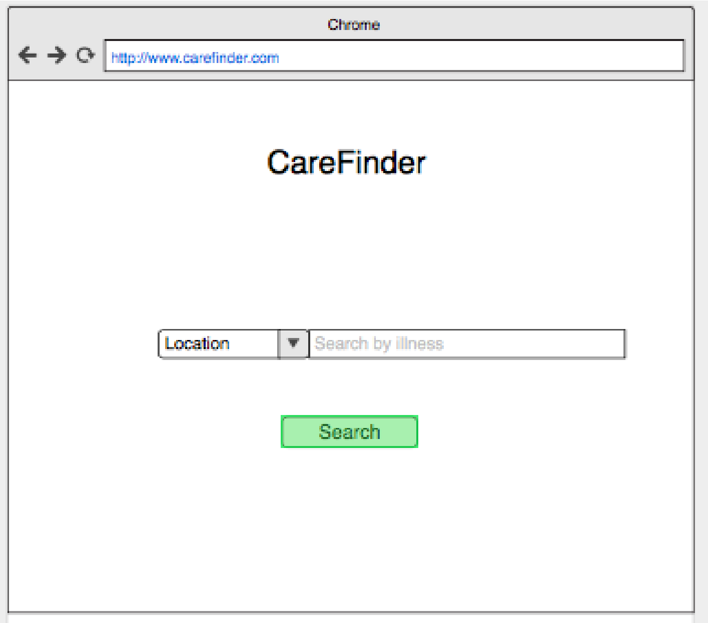
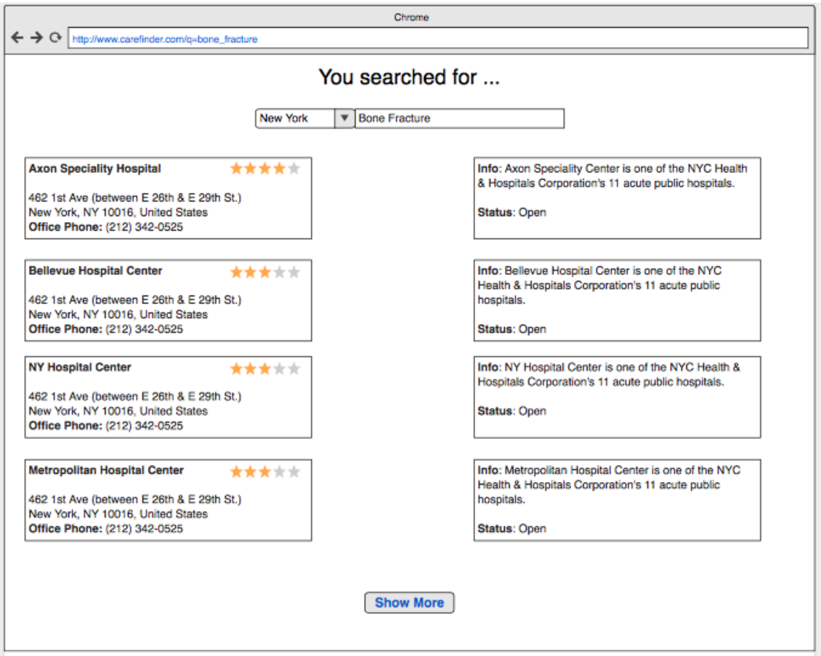


Figure 2: Interface for the Index page.

The interface for the index page is shown in figure 2. The aim is to let the users search by the *location* and the *illness*. Figure 3 shows the mockup of a basic **search engine results page** (SERP) that we have planned for *CareFinder*. The idea behind the basic interface is to display the most important information a user could possibly need on the result page itself. Valuable information related to the hospital, like, the address, contact number, general info and the status of their availability are all visible for a user, making his search decision quite easy. A future extension to the interface could be as the one shown in figure 4 where the location of hospital is also displayed on a map alongside the results.



You searched for ...

New York ▼ Bone Fracture

Axon Speciality Hospital ★★★★☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

Info: Axon Speciality Center is one of the NYC Health & Hospitals Corporation's 11 acute public hospitals.
Status: Open

Bellevue Hospital Center ★★★☆☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

Info: Bellevue Hospital Center is one of the NYC Health & Hospitals Corporation's 11 acute public hospitals.
Status: Open

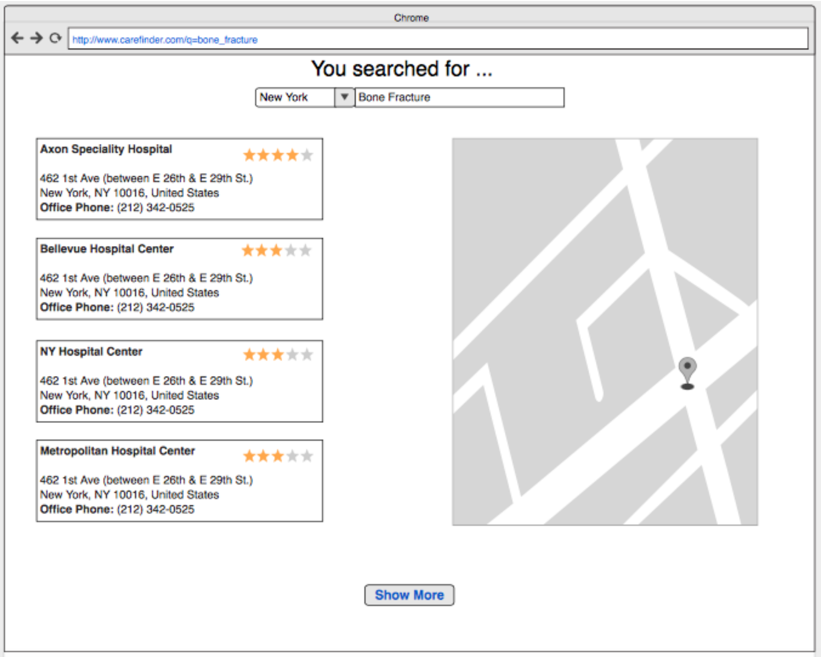
NY Hospital Center ★★★☆☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

Info: NY Hospital Center is one of the NYC Health & Hospitals Corporation's 11 acute public hospitals.
Status: Open

Metropolitan Hospital Center ★★★☆☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

Info: Metropolitan Hospital Center is one of the NYC Health & Hospitals Corporation's 11 acute public hospitals.
Status: Open

Figure 3: Basic user interface for the SERP.



You searched for ...

New York ▼ Bone Fracture

Axon Speciality Hospital ★★★★☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

Bellevue Hospital Center ★★★☆☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

NY Hospital Center ★★★☆☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

Metropolitan Hospital Center ★★★☆☆
462 1st Ave (between E 26th & E 29th St.)
New York, NY 10016, United States
Office Phone: (212) 342-0525

Figure 4: Extended interface for the SERP.

5 Execution Plan

Figure 5 shows the timeline of our project. The breakup of the timeline and the work distribution is as follows:

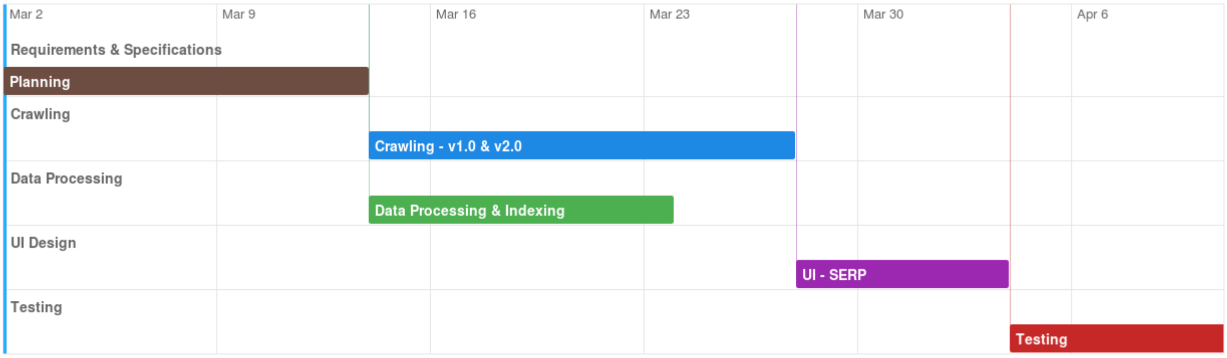


Figure 5: Timeline of our project.

- **Week 04** - Planning interface, requirements and architecture of the system - Everyone
- **Week 05** & **Week 06**
  - Location Crawler - Mingcheng
  - Hospital Processor - Wouter
  - Hospital URL Crawler - Wouter
  - Doctor Crawler - Shashank & Mingcheng
  - PubMed Crawler - Siyu & Jens
  - PubMed Processor - Siyu & Shashank
  - Clinical Trials Crawler - Jens
  - Clinical Trials Processor - Jens
- **Week 07** - Data Storage - Wouter & Jens
- **Week 07** - Build SERP - Shashank
- **Week 08** - Testing - Everyone

6 Business Plan

This project aims to help people find the hospitals or doctors based on their locations that have a good reputation or the specializations that their symptoms or diseases relate to. Once there is a large enough amount of users getting involved in this service, it will be possible to aggregate search queries, such as genders, ages, locations and diseases. Based on the aggregation of queries, we can collect the latest information about the current situations of all kinds of illnesses and we are able to make accurate predictions of different diseases activities. These information and predictions are valuable resources for medical and health care research area, which can lead to cooperations with WHO, local health organizations and health authorities.

Once our users reach the satisfying results of hospitals and doctors, it will be convenient for them if they can make an appointment through our system. A health care appointment service website is a potential business with a big market. For example, a similar website named Practo[5] from India, it has gathered $179 million funds so far. They have also extended their services to five other countries: Singapore, Malaysia, Indonesia, Philippines and Brazil. Furthermore, they have a collaboration with Uber offering their customers with door-to-door service to hospitals. This kind of project offers a modern health care service to our society with saving a lot of manpaower and material resources.

[5] https://www.practo.com/

7 Evaluation

At the end of this project there should be a working prototype. But how do we check whether our project has been a success?

Usage Evaluation
This evaluation is a subjective thing. We will evaluate whether our User Interface is friendly and intuitive enough, so that anyone can use it.

System Implementation Evaluation
This implementation can be defined into two parts. The first thing to evaluate is whether we implemented all parts of the system. That is, do we actually have a working prototype that fulfills the minimum requirements we have set.

The second thing we will evaluate is the coverage of doctors, quantity should not affect quality. The rank of a hospital that has a couple of extremely good doctors, should not be affected by a hospital that just has a sheer amount of doctors.

Relevancy Evaluation:
Finally, we will evaluate the relevancy of our system using the Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002). Testers will be asked to search for an illness using our system. Then they will have to rank the relevance of the top 10 hospitals with regards to the illness they searched for. The possible grades that can be given are are Not Relevant, Partially Relevant, Relevant or Highly Relevant. We then map these grade to a scale of 0 to 3 and can calculate the DCG from this. Finally we will compare this to the DCG of the ideal order the user has provided. We want the difference between the two to be as small as possible. With the help of crowdsourcing, we should be able to get a large testing set to compare with.

References

[1] Stuart Rose et al. "Automatic keyword extraction from individual documents". In: *Text Mining* (2010), pp. 1–20.

[2] Ian H Witten et al. "KEA: Practical automatic keyphrase extraction". In: *Proceedings of the fourth ACM conference on Digital libraries*. ACM. 1999, pp. 254–255.

[3] Qing Li, Wenhao Zhu, and Zhiguo Lu. "Predicting Abstract Keywords by Word Vectors". In: *International Conference on High Performance Computing and Applications*. Springer. 2015, pp. 185–195.

[4] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated Gain-based Evaluation of IR Techniques". In: *ACM Trans. Inf. Syst.* 20.4 (Oct. 2002), pp. 422–446. ISSN: 1046-8188. DOI: 10.1145/582415.582418. URL: http://doi.acm.org/10.1145/582415.582418.