

# Research Summary and Plan

Wu Jiang  
wujiang@ccs.neu.edu

I started to do independent research on computational intelligence when I was a junior. At that time, my major was optics and electronic engineering. I found that doing research on computer science was more interesting to me — I wanted to use algorithms to work on problems that affected daily life. A good program is like a piece of fine art. The art of problem solving, specifically in the field of computer science, can be applied to nearly every field.

Before I got my bachelor's degree at Nanjing University of Posts & Telecommunications, I authored a paper which was published at the 19<sup>th</sup> International Conference on Pattern Recognition.

I started working towards my master's degree in Computer Science at Northeastern University from Sept. 2009. During my first year I took classes in Program Design Paradigms, Algorithms, Artificial Intelligence, Theory of Computation, and Information Retrieval.

My research in Northeastern University started in May 2009.

I got to know my summer research project advisor Professor Carole Hafner<sup>1</sup> during her AI class. I continued to work with her and a classmate, Karl Wiegand<sup>2</sup>, to find a research project over the summer. Looking into the field of informatics we contacted Dr. Leonard D'Avolio<sup>3</sup> from the VA Boston Healthcare System who directed us towards the i2b2 project<sup>4</sup>. I was intrigued by the opportunities that natural language processing technologies could contribute towards helping social health care systems.

During Professor Javed Aslam's<sup>5</sup> information retrieval class, I realized the importance as well as my own interest in the subject. In Sept. 2009, I began to do research on evaluating IR systems where judgments are incomplete under Professor Aslam.

## 1 Research as a Undergraduate Student

In 2007, Frey, et.al.[4] proposed a very good clustering algorithm called “affinity propagation” in Science. Two friends and I modified the definition of “self-similarities”[5] and applied this algorithm to lossy image processing. We finished this project without any mentorship.

## 2 Research as a Graduate Student

### 2.1 Information Extraction

From May 2009 to August 2009, we did a research project on extracting medication-related information from patient health records. We were given some training data and 10 corresponding ground truth as an annotation set. Suppose our goal is to interpret this data <sup>6</sup>

---

<sup>1</sup><http://www.ccs.neu.edu/home/hafner/>

<sup>2</sup><http://www.ccs.neu.edu/home/wiegand/>

<sup>3</sup><http://www.ldavolio.org/>

<sup>4</sup><https://www.i2b2.org/NLP/Medication/>

<sup>5</sup><http://www.ccs.neu.edu/home/jaa/>

<sup>6</sup>This example is from mock data following the format of the original. The original is protected under a non-disclosure agreement.

“Patient was taking fluocinonide 0.5% cream 1 bag p.o. from Jan 12 to May 15 this year X 3 q.d. until ready for d/c home. Before this, the patient had a 50-point hematocrit drop.”

If this piece of text is on the 37<sup>th</sup> and 38<sup>th</sup> line of a patient record. The ground truth for this example is as follows.

m=“fluocinonide 0.5% cream” 37:3 37:5 || do=“1 bag” 37:6 37:7 || mo=“p.o.” 37:8 37:8 || f=“X 3 q.d.” 37:17 38:0 || du=“from Jan 12 to May 15 this year ... until ready for d/c home” 37:9 37:16 38:1 38:5 || r=“50-point hematocrit drop” 38:12 38:14 || ln=“narrative”

Where “m” means “medication name”, “do” means “dosage”, “mo” means “mode for the medication”, “f” means “frequency to take the medication”, “du” means “duration”, “r” means “reason to take the medication”, “ln” indicates “the information for this medication is from a narrative or a list”. The content within the double quote is the content for a specific field, and the number “37:3 37:5” is the offset of “fluocinonide 0.5% cream” meaning it lies in line 37, within column 3 and column 5. If nothing is found, then use “nm” (not mentioned) instead. Lines begin with one and columns begin with zero.

The challenges here are:

- There is not a standard dictionary including all the medication names.
- Terms for each field of a medication vary from record to record (doctors have different habits for term usage). Even for a single term, there are many varieties of short hand.
- There are only 10 ground truth provided, which is far too few to apply machine learning technologies.

In order to solve the first problem, we used a combination of the Orange Book and RxNorm databases, which covers 87% of the medication name from the ground truth. Most of the medication names in the ground truth files that were not found were either misspellings, abbreviations, or generic names like “antibiotics” or “pain medication”.

Because the number of the ground truth files provided is very limited, we decided to manually extract information from the rest records, five records each person per week. It turned out that these were extremely difficult to interpret manually — we spent around three weeks finishing the first five records with still hundreds remaining. I have to mention here, this is one of the main reasons that I do research on the evaluation of IR systems where judgments are *incomplete*. During this period of time, I wrote a program to color the content of different fields in these medical records based on their ground truth data in Python. This program helped us to examine how well our system worked.

We divided our task into three parts. My own was to extract a medication’s frequency. There are three basic categories: frequency, like “b.i.d”, “X 3 daily”; expressions that mean as needed, like “prn”, “as necessary”; temporal phrases that specify when a medication should be taken, like “after meal”, “at 4pm”. Also, they may be combined together, like “x 3 a day after meal as needed”.

We developed a simple algorithm which we called Medication Frequency Decision Algorithm. It is shown in Algo. [1, 2, 3].

```

Given a medication offset  $pos$  ;
Initialization:  $f \leftarrow \text{“nm”}$ ,  $l \leftarrow C_1$ ,  $r \leftarrow C_2$ ;
REsearch( $r$ ,  $pos$ );
if  $f == \text{“nm”}$  then
| REsearch( $l$ ,  $pos$ );
end
Return  $f$ ;

```

**Algorithm 1:** Medication Frequency Decision Algorithm

$C_1, C_2, \alpha$ , left string length, right string length, and span length respectively, are constant. UNITLIST contains most of the possible single element frequency strings. All these were obtained by analyzing the given

Given string  $length$ , and its starting offset  $pos$ ;

```

while not at the end of the string do
  | using regular expression searching to find all the potential individual word for frequency, store
  | their offset in a list  $OL$  [ $pos_1, pos_2, \dots$ ];
end
if  $\|OL\| == 0$  then                                     /* no word was found */
  |  $f \leftarrow$  "nm";
else if  $\|OL\| == 1$  and  $UnitCheck(word(pos_1), UNITLIST)$  then /* one-word frequency */
  |  $f \leftarrow$   $word(pos_1)$ ;
else if all  $pos_i$  are continuous or  $span(pos_i, pos_{i+1}) \leq \alpha$  then /* discrete under condition */
  | append all the words;
  |  $f \leftarrow$   $word(pos_{append})$ ;
else
  | foreach  $pos_i$  in  $OL$  do
  | | if  $UnitCheck(word(pos_i), UNITLIST)$  then
  | | | append  $f$  with  $pos_i$ ;
  | | end
  | end
end
Return  $f$ ;

```

**Algorithm 2:** REsearch Algorithm

Given a word and a list UNITLIST;

```

if  $word$  in  $UNITLIST$  then                               /* check the given word is in UNITLIST or not */
  | Return TRUE;
else
  | Return FALSE;
end

```

**Algorithm 3:** UnitCheck Algorithm

ground truth data and extracted data manually. We found the best result when  $C_1 = 5, C_2 = 20, \alpha = 2$ . Although we have not received results for larger sets of testing data, our algorithm was very effective for the training set.

## 2.2 Information Retrieval

When doing the information extraction project, I was also taking Professor Javed Aslam's information retrieval class. From the summer research project, I felt that it was very painful to interpret data manually from documents. I found that Aslam was doing some very interesting research on IR systems evaluation where judgments are incomplete.

From Sept. 2009, I have been working on IR evaluation under Professor Javed Aslam.

Buckley, et.al.[1] showed that average precision (AP) is not robust enough for incomplete relevance judgments. They proposed a measure *bpref*, unfortunately it was without strong theoretical support. Not soon after, Yilmaz, et.al. completed research on how to estimate average when judgments are incomplete, both when sampling is uniformly random[7] as well as when weighted[8]. They proposed several better methods: *indAP*, *subAP*, *infAP*, and *xinfAP*. All the methods are based on an assumption that AP is the "gold standard" for evaluation. On the other hand, nDCG (Discounted Cumulative Gain) reflected other features of IR systems. According to Croft, et.al.[3], the nDCG's formula has no theoretical support at all. We are trying to find some connections between AP and nDCG.

Sampling strategy is very important for evaluation IR systems. Three interesting papers that I am reading are Aslam et.al.'s "A Practical Sampling Strategy for Efficient Retrieval Evaluation" (draft), Pavlu's PhD thesis "Large Scale IR Evaluation"[6] and Carterette et.al.'s "If I Had a Million Queries"[2].

## 2.3 Algorithms and Complexity

I am also taking a topic course — Algorithmic power tools, this semester taught by Professor Rajmohan Rajaraman<sup>7</sup>. At the first half semester, we mainly focus on studying *Approximation Algorithms, Probabilistic Method, Lovasz Local Lemma, Entropy Compression Argument, Randomized/Deterministic Rounding, Linear Programming, Generalized Network Design*. The next half semester, I will do a research project under the topic of "Algorithms and complexity of periodic scheduling". Papers I need to read include Eisenbrand, et.al.'s "EDF-schedulability of synchronous periodic task systems is coNP-hard" and Bonifaci, et.al.'s "Algorithms and Complexity for Periodic Real-Time Scheduling".

## 3 Research Internship Plan

Although life does not always go as planned, it is always helpful to think ahead. The following would be my plan for my research internship.

- Explore further in machine learning and information retrieval research.
- Contribute what I have learned on algorithms, machine learning to the group.
- Organize a weekly seminar in theoretical information retrieval if one does not currently exist.
- Get a lot of surprises.

## 4 Further Plan in Five Years

After coming back from the internship, I will be spending one semester to finish my master's degree, then I hope to begin my PhD on information retrieval, machine learning, and information theory.

---

<sup>7</sup><http://www.ccs.neu.edu/home/rraj/>

## References

- [1] Chris Buckley and Ellen M. Voorhees, *Retrieval evaluation with incomplete information*, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, July 2004, pp. 25–32.
- [2] Ben Carterette, Virgil Pavlu, Evangelos Kanoulas, Javed A. Aslam, and James Allan, *If I had a million queries*, Advances in Information Retrieval: 31st European Conference on IR Research (Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, eds.), Lecture Notes in Computer Science, vol. 5478, Springer-Verlag, April 2009, pp. 288–300.
- [3] Bruce Croft, Donald Metzler, and Trevor Strohman, *Search engines: Information retrieval in practice*, Addison-Wesley, 2009.
- [4] Brendan J. Frey and Delbert Dueck, *Clustering by passing messages between data points*, Science **315** (2007), 972–976.
- [5] Wu Jiang, Fei Ding, and Qiaoliang Xiang, *An affinity propagation based method for vector quantization codebook design*, ICPR08, December 2008, pp. 1–4.
- [6] Virgil Pavlu, *Large scale IR evaluation*, Ph.D. thesis, Northeastern University, 2009.
- [7] Emine Yilmaz and Javed A. Aslam, *Estimating average precision when judgments are incomplete*, Knowledge and Information Systems **16** (2008), no. 2, 173–211.
- [8] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam, *A simple and efficient sampling method for estimating AP and NDCG*, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, eds.), ACM Press, July 2008, pp. 603–610.