Extracting Novel Ideas From Crowdsourcing Requirements

Diem Dao
Advisor: Dr. Pradeep Murukannaiah

R·I·T
B. Thomas Golisano College of Computing and Information Sciences

Background

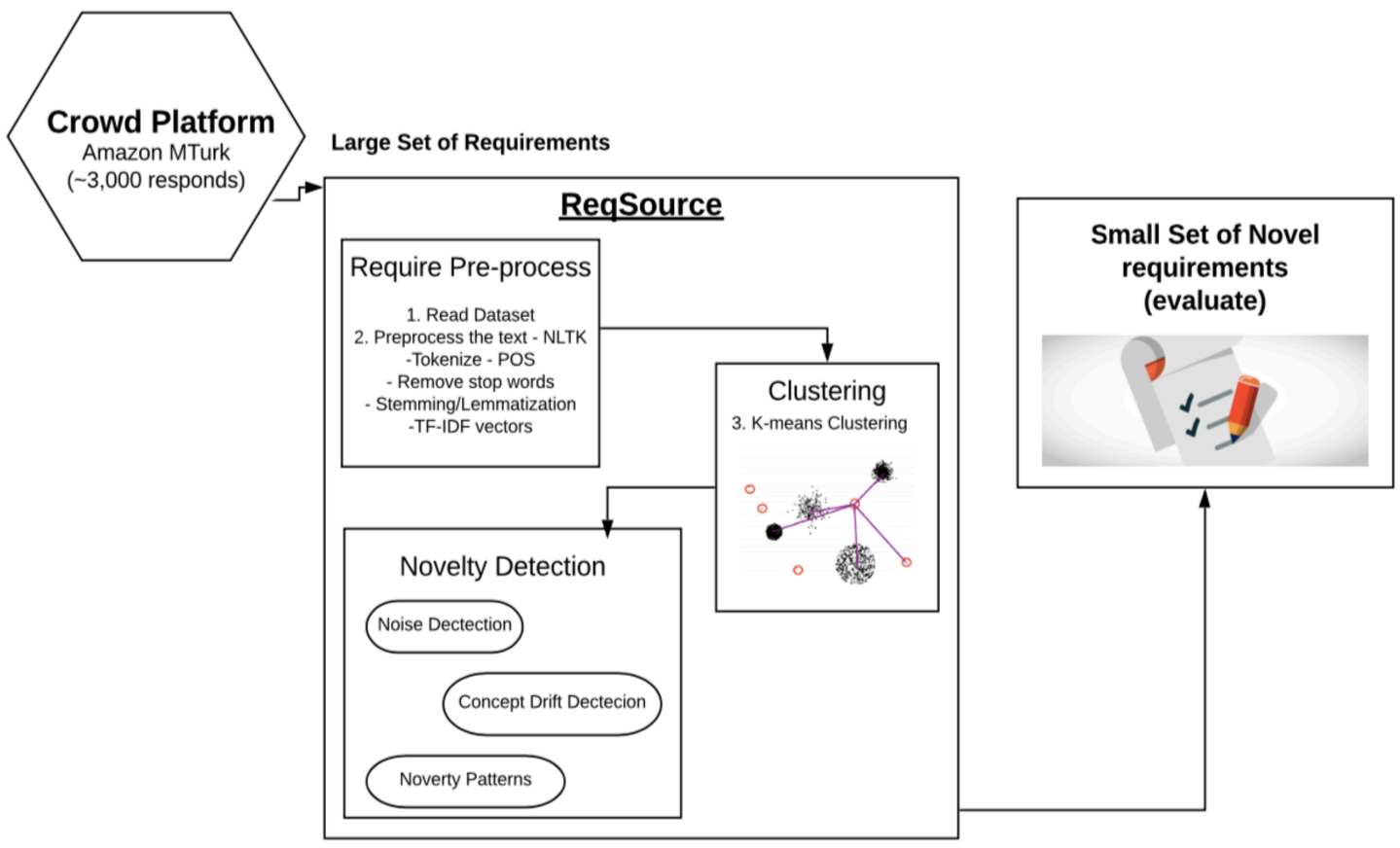
- When researchers and businesses crowdsource data from the public, a large database of information can be overwhelming and redundant.
- Finding novelty ideas from a large data set can be challenging.

Hence, we developed a study that:

- *Can automatically determine novel ideas from crowdsourcing requirements.*
- *Collected approximately 3,000 requirements from a MTurk survey on what people wanted in their smart home applications.*
- *Filtered the mundane ideas from novel suggestions (requirements).*

Research Design

Goal: Develop tools and techniques to extract novel ideas and requirements from crowdsourced data.



Current Progress

- Carried out data pre-processing techniques using NLTK (Tokenization, Stop word removal, POS Tagging, Stemming from textual requirements obtained through crowdsourcing).
- Used data mining techniques such as TF-IDF Vectorization and Count Vectorization to understand and analyze the data.
- Implemented K means algorithm and used the Elbow method with Euclidean distance to determine the optimal number of clusters.
- Find the requirements in each cluster.

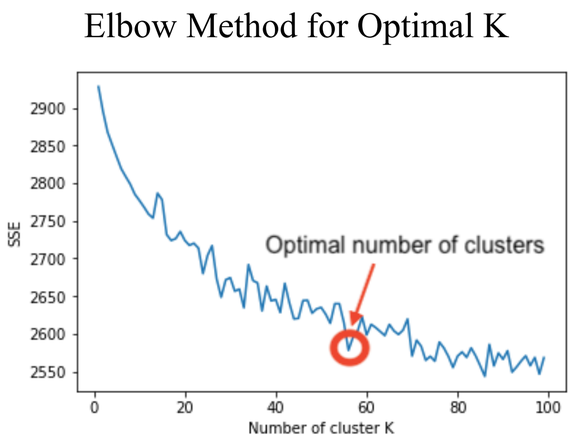


Fig. 1 shows the optimal number of K clusters

```
Top terms per cluster:
Cluster 0: applianc turn breakfast use power
Cluster 1: oven preheat turn cook time
Cluster 2: door open garag unlock front
Cluster 3: vacuum cleaner automat floor clean
Cluster 4: pet food water automat door
Cluster 5: water plant heater hot automat
Cluster 6: know let left run leav
Cluster 7: fridg freezer low open know
Cluster 8: my temperatur bath adjust automat
Cluster 9: motion sensor detector outdoor light
Cluster 10: music play sooth shower room
Cluster 11: lock window cabinet unlock refridger
Cluster 12: automat turn wall time toilet
Cluster 13: control heat floor temperatur cold
Cluster 14: light turn adjust dim day
Cluster 15: video game monitor record outsid
Cluster 16: alarm clock sound set system
Cluster 17: phone via alert text mobil
Cluster 18: window automat open sensor close
Cluster 19: clean self toilet automat devic
Cluster 20: remind medicin homework bill take
Cluster 21: washer dryer cloth start load
Cluster 22: air condition qualiti purifi condit
Cluster 23: energi use usag save consumpt
Cluster 24: temperatur adjust room automat outsid
Cluster 25: kitchen applianc recip control built
Cluster 26: electron abl remot devic turn
Cluster 27: keep track health pet my
Cluster 28: room light turn leav enter
Cluster 29: autom system door sprinkler music
Cluster 30: blind close open get curtain
Cluster 31: comput someon use turn phone
Cluster 32: shower voic activ command certain
Cluster 33: go get sleep light know
Cluster 34: coffe machin pot maker am
Cluster 35: cloth dryer closet dri wash
Cluster 36: dog alert collar bowl door
Cluster 37: monitor health babi room level
Cluster 38: kid children school alert get
Cluster 39: system secur sound entertain room
Cluster 40: show favorit tv televis record
Cluster 41: alert someon mail time everi
Cluster 42: lock door automat front unlock
Cluster 43: food tell low fridg dispens
Cluster 44: stove automat cook oven ga
Cluster 45: tv turn channel get room
Cluster 46: fire case turn automat depart
Cluster 47: notifi someon cat leak my
Cluster 48: sensor detect child moistur tell
Cluster 49: detector smoke carbon monoxid ga
Cluster 50: solar panel roof power adjust
Cluster 51: thermostat room temperatur adjust learn
Cluster 52: bed sleep make warm pattern
Cluster 53: wash machin need load run
Cluster 54: light automat turn dark room
```

Fig. 2 shows the 55 most common ideas and concepts from all the requirements

	cluster	data_index	text
4	1	4	class, remind
152	1	152	remind, pre, manag, list, day, night
516	1	516	remind, activ, kid, need, go
773	1	773	set, remind, start, laundri, alert, laundri, done
799	1	799	make, reccomend, date, new, movi, may, enjoy, ...
848	1	848	remind, day, everi, morn, get
879	1	879	remind, leav, left, food, water, cat
987	1	987	remind, exercis, track, progress
1010	1	1010	A, bill, remind
1011	1	1011	auto, featur, remind, chang, water, sofner
1099	1	1099	remind, homework

Fig. 3 shows what requirements are in each cluster (subset of a group). This cluster shows concepts relevant to the topic of reminders.

Technologies

Direction

- Implement MINAS: multiclass learning algorithm for novelty detection in data streams.
- Create a decision model by training and testing the data set to detect novelty patterns.
- Remove outliers or noise and find concept drifting.

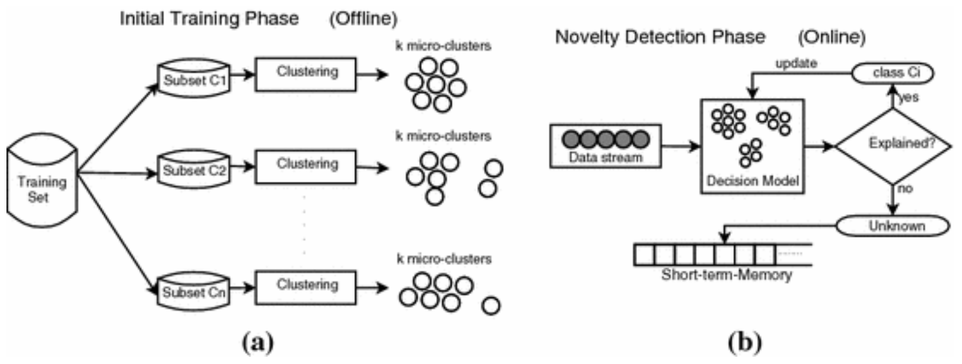


Fig. 4 MINAS algorithm in 2 phases: Initial Training Phase (Offline) and Novelty Detection Phase (Online)

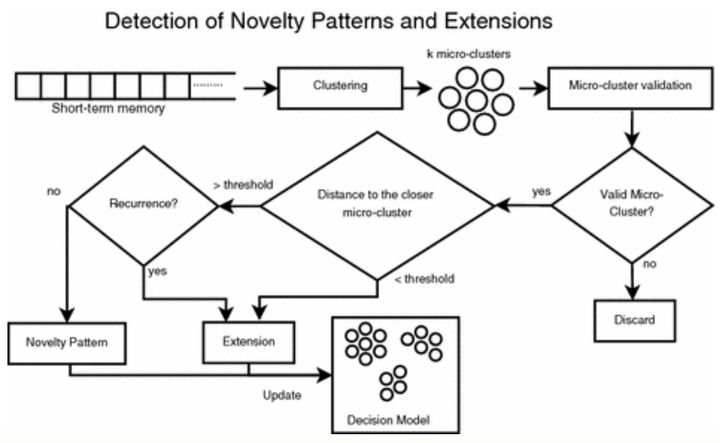


Fig. 5 Novelty Detection Process

References:

de Faria, E.R., Ponce de Leon Ferreira Carvalho, A.C. & Gama, J. Data Min Knowl Disc (2016) 30: 640. https://doi.org/10.1007/s10618-015-0433-y

Acknowledgement

This work is supported by NSF CNS-1757680.

Contact: Pradeep K. Murukannaiah, Ph.D.
Assistant Professor
Department of Software Engineering
Rochester Institute of Technology
pkmvse@rit.edu