



Prereqir: Recovering Pre-Requirements via Cluster Analysis

*Jane Huffman Hayes, Giuliano (Giulio) Antoniol and
Yann-Gaël Guéhéneuc*

Content

- Problem Statement
- PREREQUIR Idea
- PREREQUIR Process
- Technologies
- WEB Browser Requirements
- Case Study Results
- Conclusions

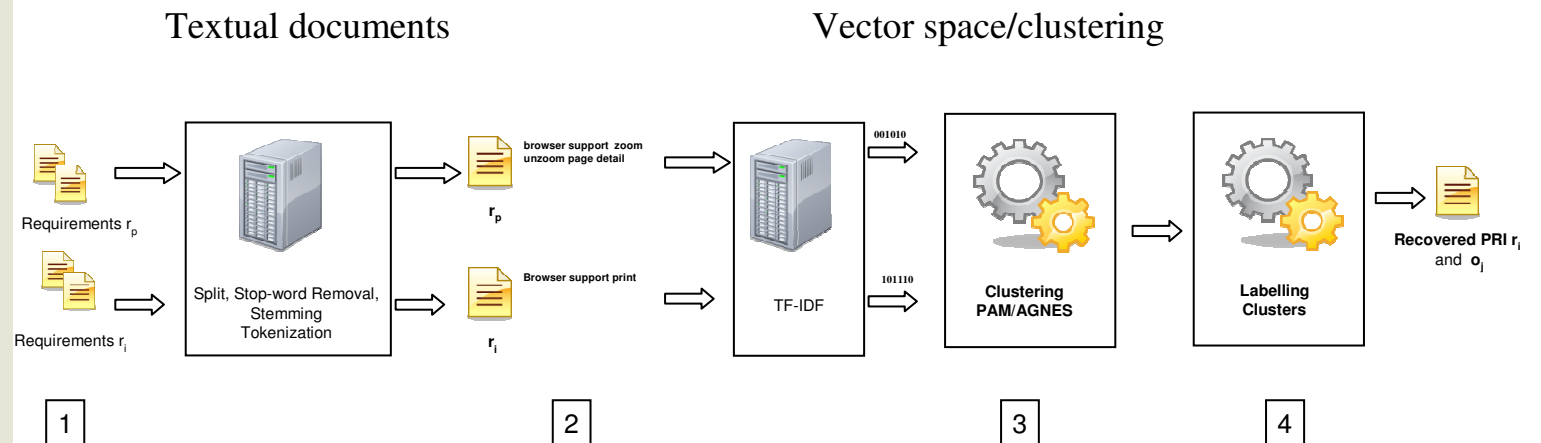
The Challenge

- A few years after deployment, the RS may no longer exist.
- If it exists, it will be almost surely outdated.
- My customers may desire new functionalities or technologies that my system may or may not implement.
- I poll my stakeholders:
 - programmers, managers, testing team members, marketing personnel, and end users;
 - find out what they believe the system should do.

PREREQIR in Essence

- We need a pre-requirement document:
 - what the competitor systems do;
 - what our customer base needs.
- Obtain and vet a list of requirements from diverse stakeholders.
- Structure requirements by mapping them into representation suitable for grouping via pattern-recognition and similarity-based clustering.
- Analyze clustered requirements to divide them into set of essential and set of optional requirements.

The PREREQUIR Process



PREREQUIR Technology

- Standard information retrieval vector space model.
- Indexing process:
 - Stopper;
 - Stemmer;
 - Thesaurus (not vital but helps);
 - TF-IDF indexing.
- Clustering PAM and AGNES.
- Labeling: still an open question.

Step 1 – Collect Stakeholders RS

- By means of questionnaires, collect stakeholders requirements.
- We favor a non-intrusive lightweight approach such as a WEB based questionnaire.
- Minimize the risk of influencing stakeholder.
- There is risk that:
 - he/she did not really understand the task;
 - the granularity and level is very different between respondents;
 - the respondent population is not heterogeneous enough;
 - the sample size is small.

Step 2 – Vector Space Mapping

- The goal is to group single requirements by different users into clusters representing the same functionality/concept.
- By means of standard IR tools, map the collected requirements into a vector space.
- Stopper, stemmer, and TDF/IDF plus thesaurus expansions:
 - certain stakeholders may use cryptic terms such as RFC or test/benchmark acronyms.

Step 3 – Clustering

- Transform similarity into a distance.
- Apply robust partition around medoids.
- Estimate the number of clusters (different requirements) *silhouette*:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

>0.70 very strong structure
0.50 ... 0.70 reasonable structure
0.25 ... 0.50 weak structure
< 0.25 no structure.

- $a(i)$ average distance to the other PRI in the cluster;
- $b(i)$ is the average distance to PRI in the nearest cluster.

- Take the flex close to max value of the average silhouette.

Step 3 Bis – Tree Structure

- If there is a weak structure, check for a requirement tree organization.
- Re-cluster with AGNES.
- Compute the Agglomerative Coefficient (AC).
- AC measures the strength of the hierarchical structure discovered.
- $AC > 0.9$ a very strong hierarchical structure.
- Impose a threshold on the average similarity to avoid grouping “too different” things.

Step 4 – Label Clusters

- Process each PRI of a cluster:
 - stopping, stemming;
 - build cluster-specific dictionary;
 - weight each word by its frequency in the cluster:
 - If a word is in all the PRI in a cluster, its weight is 1.00. If a word appears in half of the PRI, its weight is 0.50.
- For a given stemmed PRI, calculate a score:
 - sum up the weights of the stems present in the cluster dictionary to obtain a *positive* weight;
 - count the number of words in the cluster-specific dictionary that are absent in the current PRI:
 - obtain a *negative* weight.
- Assign a score to the PRI computed as:
 - the ratio positive weight / negative weights.
- Label the cluster:
 - take the PRI with the highest score.

Case Study

- Mimic the recovery process for a Web browser.
- Pool via e-mail to a set of users (about 200).
- 25 answers out of which we kept 22, overall 433 user needs:
 - mostly male (20), age varies, average 36, standard deviation 9.5;
 - respondents: 10 researchers, five lecturers/professors, four students, one programmer, and two project managers.

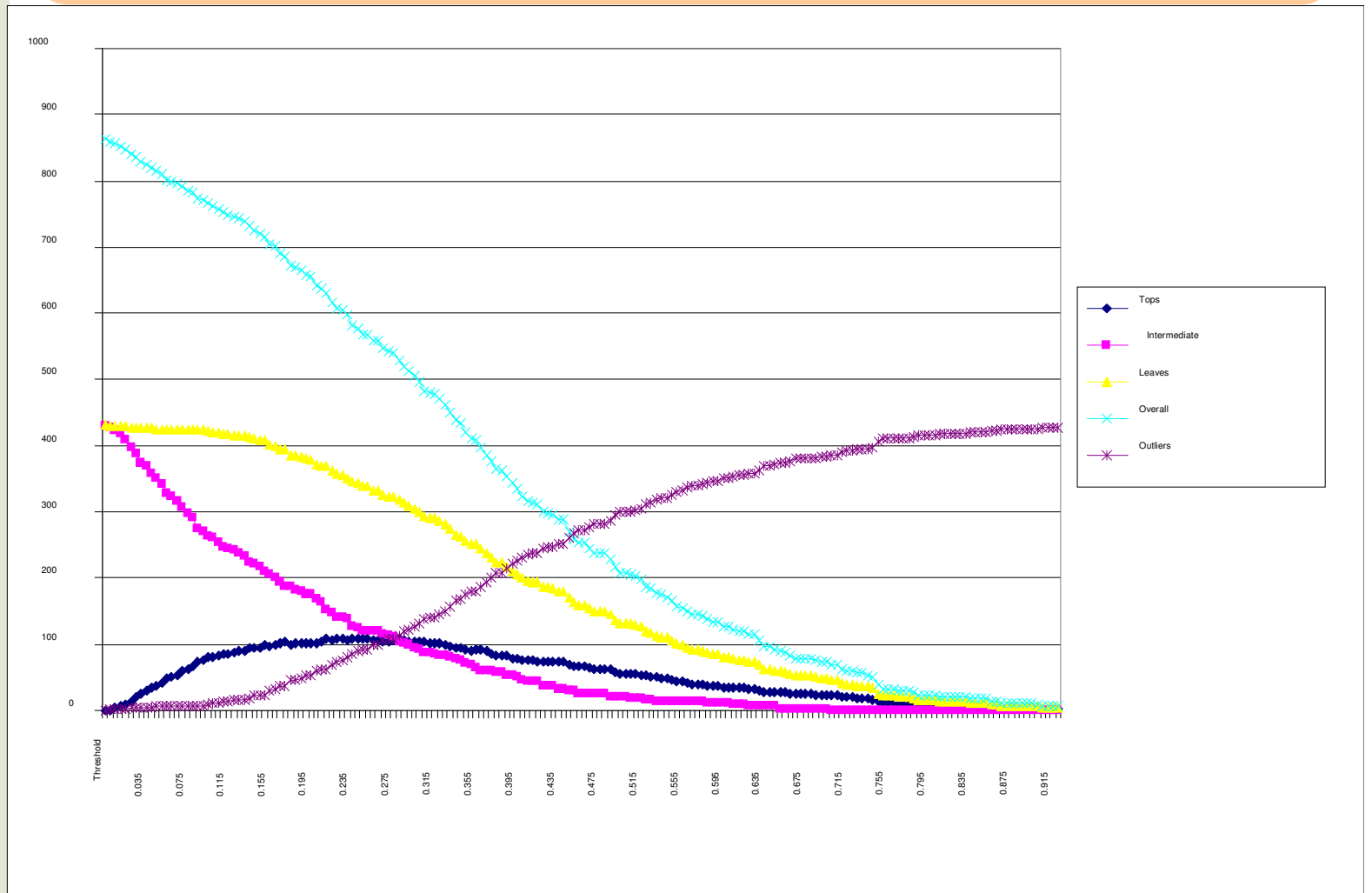
PAM - AGNES

- We did not find a strong or evident cluster structure:
 - silhouette about 0.26;
 - region between 167 – 170 cluster:
 - say 170 clusters or less.
- AGNES reports a strong structure:
 - AC above 0.9.
- Grouping via AGNES
 - grows a tree starting from leaves

Outliers

- Setting a cluster internal similarity threshold decides
 - top level clusters
 - singleton clusters - outliers
 - inner nodes
- The “non kept” are also important:
 - single user needs;
 - more expert users may use acronyms
 - must comply with ACID2
 - “too generic: sentences:
 - it should be fast.

AGNES Clusters

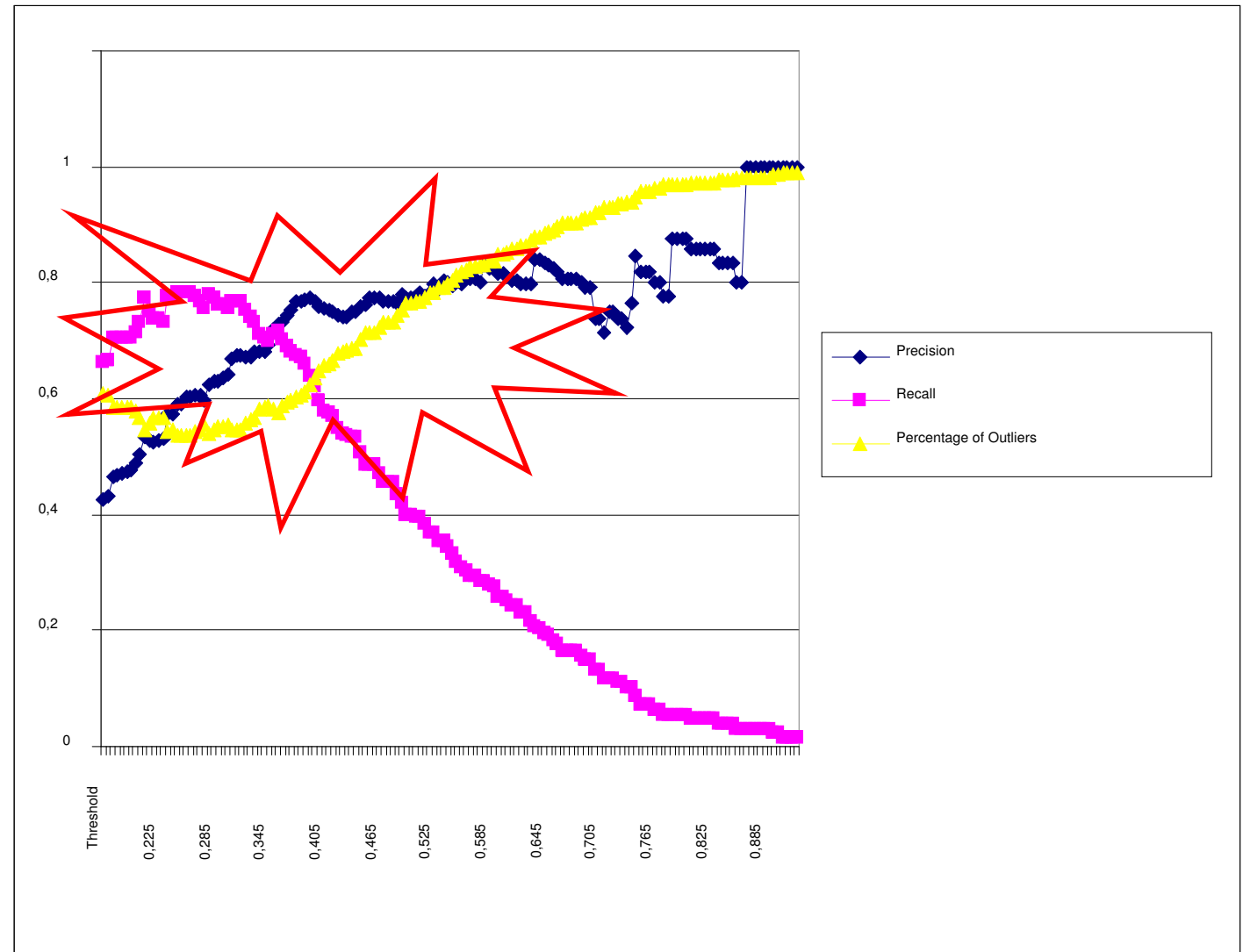


Manual Verification

- Two people reviewed cluster and cluster labeling.
- IR measures precision and recall.
- Precision measures the quality of the clusters.
- A conservative approach:
 - “Yes” was assigned if both authors said “Yes”;
 - “No” was assigned if one of the authors said “No”;
 - “Maybe” was assigned in the other cases.

Precision Recall – 0.36

128 Common User Needs, 181 Outliers



Traceability Task

- PRI for a Web browser provided:
 - Web site: www.learnthenet.com.
- There are 20 LtN PRI:
 - textual PRI ranging from 5 to 73 words, having on average 23.5 words.
- LtN10: “The toolbar should include a Reload or Refresh button to load the web page again.”
- Trace via vector space retrieval with *tf-idf*.
- Similarity threshold of 0.20.

Manual Evaluation by Two Authors

- 14 of the 20 LtN PRI are traced:
 - the 14 PRI were all marked as “Yes” by both authors.
- If we also include the two marked as “Maybe” there are 16 LtN PRI out of 20 traced.
- Overall, between 70% (“Yes” only) and 80% (“Yes” and “Maybe”) of the LtN PRI are also found in the PRI obtained from the respondents.

Threats to Validity

- *External validity*: only one system and 22 answers out of 200, impact of vocabulary is not known.
- *Construct validity*: computation performed using widely adopted toolsets, other tool can produce different results.
- *Reliability validity*: material will be made available.
- *Internal validity*: subjectivity introduced by experts, “Yes” if and only if both agrees.

Conclusion

- AGNES clusters PRI with an accuracy of 70%.
- A similarity threshold of about 0.36, about 55% of the PRI were common to two or more stakeholders and 42% were outliers:
 - 128 – 181.
- We automatically label the common and outlier PRI with 82% of the labels being correct.
- The method achieves roughly 70% recall and 70% precision when compared to a ground truth.