

# Information Extraction Technique

This document describes the instructions for using the Information Extraction Technique. The document is divided into 5 main sections:

- 1) An introduction to the technique
- 2) A glossary where the key terms are defined in this context
- 3) A procedure of steps that describes clearly WHAT you have to do.
- 4) A set of guidelines, i.e. heuristics to keep in mind as you do the above procedure.
- 5) A set of data entry forms

## 1) Introduction

The Information Extraction Technique is a structured reading method for extracting information from papers, which can later be analyzed to explore the evidence that supports various hypotheses. In this assignment, you will follow the procedure for identifying and recording hypotheses from papers in the scientific literature. The procedure focuses the search specifically on hypotheses and context descriptions, providing some guidelines to help recognize and abstract them.

## 2) Glossary

### Hypothesis

A hypothesis is a tentative explanation for certain behaviors, phenomena, or events that have occurred or will occur. A good hypothesis states as clearly and concisely as possible the expected relationship (or difference) between two or more variables and defines those variables in operational, measurable terms. Any hypothesis should be stated in such a way that data can be collected that either supports or refutes the hypothesis. For the purpose of our analysis, we classify hypotheses as tested or untested:

#### a) Tested Hypothesis

A tested hypothesis is a tentative explanation for certain behaviors, phenomena, or events that have occurred in experience or empirical study.

#### b) Untested Hypothesis (Belief)

An untested hypothesis (otherwise called a belief or assumption) is a tentative explanation for certain behaviors, phenomena, or events without explicit reference to empirical data.

## 3) Procedure

- 1) Read the paper, keeping in mind the two kinds of information that you want to identify:
  - a. Hypotheses (tested and untested), and
  - b. Context descriptions
- 2) When you find relevant information during your reading, highlight it so that there can be some traceability back to the original source if questions arise later.
- 3) Transfer the key details to the data entry forms. For complete descriptions of the fields you should complete, see section 5 on “data entry forms.”

## 4) Some Guidelines

There will be at least one **context description** associated with each source (possibly more, if the paper describes data that was collected from several projects). Our experience is that the context descriptions usually come shortly after the introduction. As different studies report different metrics of interest to them, not every paper will have all of the required information for our template. However, the template should be filled out as completely as possible given the information that has been published.

Our experience is that the section of the paper describing the analysis of the empirical data is where most tested **hypotheses** can be found. The conclusions are a good place to find

hypotheses, although these are many times repetitions from earlier in the text. Some hypotheses will also be found in tables and figures; although not explicitly stated in the text of the paper, relationships that are expressed visually for readers will need to be translated into textual form to be inserted into the Hypothesis Form in a usable way.

## 5) Data Entry Forms

An excel file contains worksheets for recording:

- Context descriptions
- Hypotheses

### 5.1) Context Descriptions

Fill out one form for each paper (or each study recorded in the paper, if there are multiple). The attributes of the context description form are:

- **Paper Title:**
  - The title of the paper from which you are extracting the information.
- **Complete Reference:**
  - A complete bibliographic reference to this paper.
- **Topic:**
  - Use the IEEE keywords at the following site to denote topic categories.
    - <http://www.computer.org/mc/keywords/software.htm>
  - Choose the keywords that best describe the subject of the study described in the paper.
- **Goals:**
  - You should fill in a set of GQM goal templates for the study (using the form: Analyze O for the purpose of X with respect to M from the point of view of P in the context of C). Remember to make clear the entities being studied, (i.e., the process, product, model, metric, ...), the attributes of the entities that are of interest, the purpose of the study, (i.e., whether the study is aimed at characterizing, understanding, evaluating, predicting, or improving), and for whom the study should be of value, (i.e., a researcher, project manager, corporation,...).
- **Variables**
  - Describe as many as possible of the following characteristics for each dependent and independent variable in the study:
    - *Name*: How the variable is referred to in the paper.
    - *Possible Values*: The possible values for the variable, if controlled.
    - *Data Collection Details*: Details of the method used to measure the variable, including for example what instrumentation and tool support were used.
- **Subjects**
  - Describe as many as possible of the following characteristics for the subjects in the study:
    - *Category*: A generalized description of the experience level of subjects. Possible values here can be:
      - Undergraduate Students;
      - Graduate Students;
      - Students: This is an "unknown" type of students.
      - Professionals;
      - Scientists;
      - Unknown: Not described in the paper.
    - *Number*: The number of subjects that participated in the experiment.
    - *Incentives*: Describes the subject recruitment rewards. The possible values are
      - Grades: If students' grades were affected by their participation
      - Extra Credit: If participating students received extra credit in the class
      - Payment: The subjects were paid to take part in the experiment.
      - Other Rewards: Please specify.

- No Rewards
  - Unknown
- **Task**
  - *Category:* Categorize the tasks given to subjects according to the **tasks** applied and the **work products** they were applied to (e.g. *created a design document*).
    - Possible values of tasks:
      - Plan
      - Create
      - Modify
      - Analyze
    - Possible work products:
      - Requirements
      - Architecture/design
      - Code
      - Etc.
  - *Duration:*
    - Duration of the time that subjects had available for the task(s).
  - *Work Mode:* Select whether subjects performed the task(s) as:
    - Team
    - Individual
  - *Application Type:* The type of application on which the tasks were performed. The possible values are:
    - Constructed: Applications constructed for the purpose of the experiment;
    - Commercial: A commercial application;
    - Student Project: An application constructed for a class assignment;
    - Open Source: An open source application;
    - Unclear
- **Environment**
  - *Location:* Describe the location in which the study was run. Possible values:
    - Industry
    - Classroom
    - Other (Please specify)
    - Unclear
- **Replication:** Indicate whether this study is a replication of another one. (Choose “yes” or “no.”) Include a reference to the original experiment if this is a replication.
- **Other**
  - Note any other information that is important for understanding the model, metric, techniques, or the empirical study itself (e.g., missing definitions, environmental characteristics, or information about process conformance).

## 5.2) Hypotheses

Fill out one form for each hypothesis you identify. The attributes of the hypothesis form are:

- **Plain Text:**
  - Try to write the hypothesis using the words from the paper so that traceability is assured. When identifying hypotheses, recall that: The hypothesis should be stated in such a way that data can be collected that either supports or refutes the hypothesis. A good hypothesis states as clearly and concisely as possible the expected relationship (or difference) between dependent and independent variables and defines those variables in operational, measurable terms.
- **Type:**
  - Tested hypothesis
  - Untested hypothesis
  - Hypothesis from another paper (A hypothesis that was tested in another paper and reported in this one.)
- **Level of Support:** The support should be described as one of the following levels:

- Significantly positive: The results support the hypothesis and the results of a test are included to show that the results are statistically significant, that is, with a high degree of certainty are not the result of pure chance.
- Positive: The results in the paper support the hypothesis, but no significant statistical results can back this up.
- Null: The hypothesis has been tested but the results in the paper neither support nor contradict the hypothesis.
- Negative: The results in the paper contradict the hypothesis, but no significant statistical results can back this up.
- Significantly negative: The results contradict the hypothesis and the results of a test are included to show that the results are statistically significant, that is, with a high degree of certainty are not the result of pure chance.
- Belief: The hypothesis is formulated based on assumption or belief but has not been tested.
- **Confidence in support:** Reflects the level of confidence that can be placed in the results and author's analysis. The rating is subjective, based on your assessment of the measurements reported:
  - High: Results were rigorously measured.
  - Med: Measurement was not completely rigorous *or* the context was not realistic.
  - Low: Measurement was not completely rigorous *and* the context was not realistic.
  - None: No evidence was presented. (The hypothesis describes a belief and has not been tested.)
- **Observations:**
  - This is a free-text field for you to keep track of any additional information that is important for correctly understanding or interpreting the hypothesis.