INTROSPECTIVE SELF-EXPLANATION IN ANALYTICAL AGENTS

Paper 5

Anita Raja & Ashok Goel

2009

OVERVIEW

A relatively recent paper (2009)

Focused on the designing of analytical agents so that they can explain their analysis and conclusions to human users

Presents a suggested architecture for an agent system capable of self-introspection, and argues why it might be a good thing

ANALYTICAL AGENTS?

Automatic autonomous agents performing analysis

- Obvious use-case is decision support systems
- Paper focuses on usage in highly complex domains (such as medicine or financial analysis)

The analysis process performed by the agents:

- Recognizing and describing a problem based on data about events
- Making explaining hypotheses based on the data
- Collecting and processing additional data
- Evaluating the hypotheses
- Selecting the best hypothesis

ABDUCTIVE REASONING

“Inference to the best explanation“

Form of logical method to construct a hypothesis based on observations as input, with the goal of explaining the observations

Due to being based on uncertain observation does not guarantee the correctness of the conclusion.

INTROSPECTIVE SELF-EXPLANATION

Operates with two types of explanations: Abductive explanations and Self-explanations

Abductive explanations focus on explaining patterns in data

Self-explanations explain the work-process of the agent

- Is usually decomposed to justifying choice of data, explaining conclusions, and justifying decisions

Introspective self-explanations is when you ease the generations of the self explanations by introspecting/analyzing the decision making process

- They chose to use their own “introspective task structure”

PRIMARY APPLICATION OF THE SYSTEM

Supporting financial decision making

Generating explanations on different abstraction levels/adapting them to the user and the users goals

Abstraction↓/Analyst→	A	OA	HASO	OASO	HADO
Conclusion(s)	X	X	X	X	X
Confidence Values of Conclusions	X	X	X	X	
Justifications for Conclusions	X	X	X	X	X
Alternate hypotheses/justifications	X	X	X		X
Analyst's Assumptions and Biases	X	X	X	X	X
High-Level Explanation	X	X	X	X	
Mid-Level Explanation	X	X		X	
Low-Level Explanation	X	X			
Supporting Raw data	X				
Justifications for Domain Knowledge	X				

Table 1. Decision Process Explanation Matrix A: Analyst; OA: Other Analysts; HASO: Higher Authority, Same Organization; OASO: Other Analysts, Same Organization; HADO: Higher Authority, Different Organizations

ANALYSTS PROBLEM SOLVING PROCESS

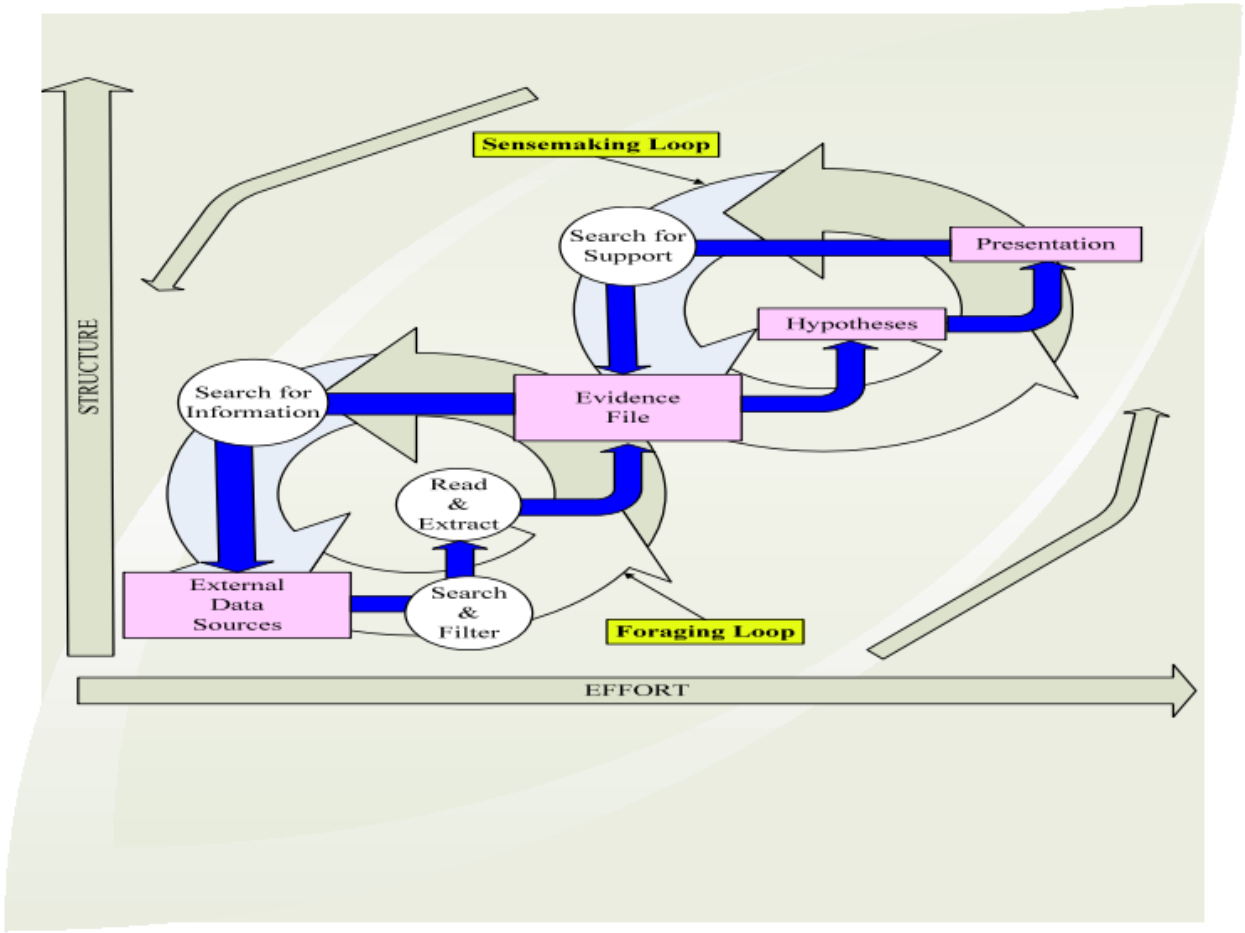

Fig. 1. Pirolli and Card's Sensemaking and Foraging Loops

MAIN SHORTCOMINGS OF HUMAN-GENERATED ANALYSES IN THE AREA

Due to bad storage solutions, humans have trouble keeping track of many explanations for prolonged periods (such as those spent on making analyses)

They tend to pick favorites fast, and mostly don't re-evaluate or update the explanations in light of new data

In addition they tend to be confirmation-biased, searching for supporting data and overlooking refuting data

SENSEMAKING SYSTEM

The sensemaking system tries to address those shortcomings

At an abstract level, what the system does is:

- Data from a present situation is tested against a collection of past cases and data
- The old stuff that matches the current data is retrieved and invoked, thereby generating hypotheses
- Hypotheses explaining different pieces of the data are combined into multiple stories that explain large parts of the current data

ILLUSTRATION OF HOW THE SENSEMAKING COMPONENT (AKA STORY ABDUCTION [STAB]) WORKS

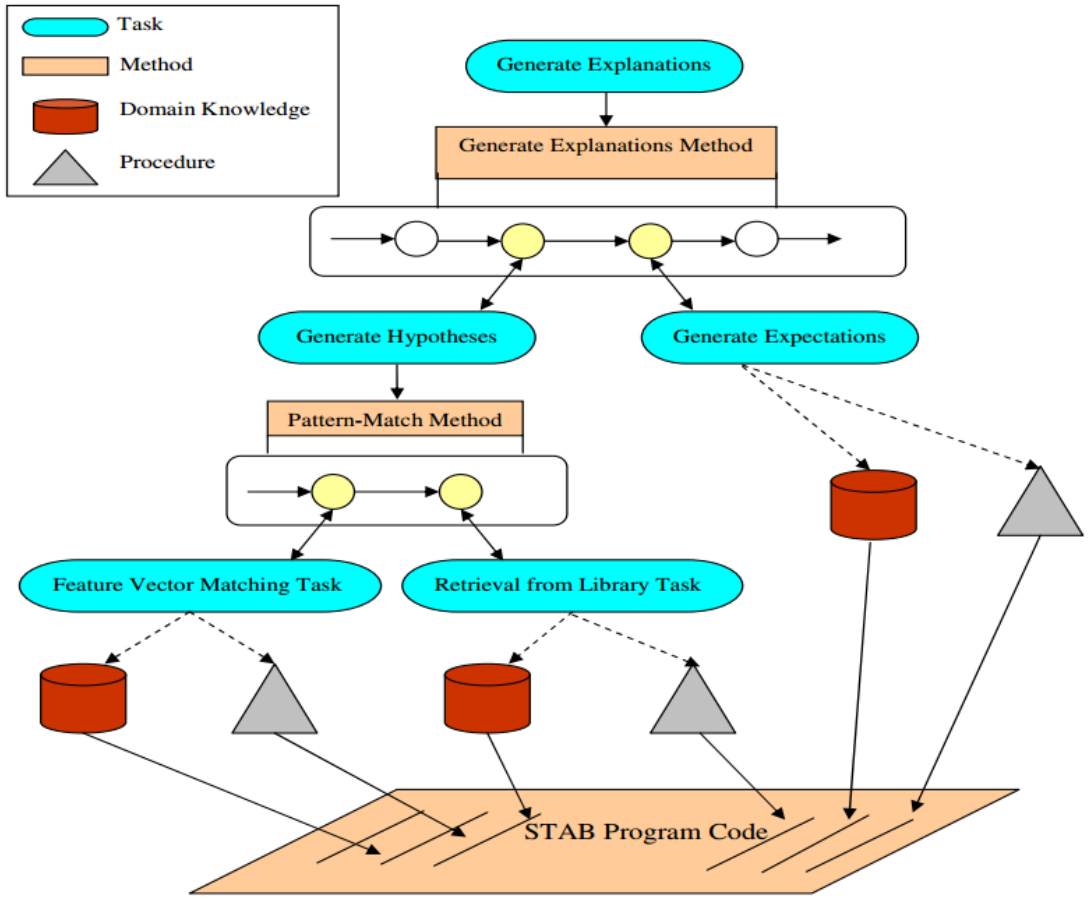

Fig. 5. Meta-STAB's Introspective Task Structure

THE FORAGING COMPONENT

TIme BOunded Reasoning (TIBOR)

Is designed to handle gathering of all kinds of types of data from many sources

Can determine how to extract data from the sources

Can also find good way to visualize the data (interactively they boast)

Supports hypothesis tracking and validation

Works in highly uncertain domains

SUMMARY

~~Appears to be a vague presentation of an explanation based CBR system, with some illustrations on how they made the different components they made work~~

~~They seemed pleased with the results, but did not indicate how (or whether) they tested them~~

It is an explanation based CBR system that is not fully implemented at this point in time

The authors however seem rather pleased with the work so far, they find that agents capable of self-introspection have a place in decision making processes, and want to continue their implementation