Exploiting Evidence Analysis In Plan Recognition[⋆]

Sandra Carberry[1] and Stephanie Elzer[2]

[1] Dept. of Computer Science, University of Delaware, Newark, DE 19716 USA
carberry@cis.udel.edu
[2] Dept. of Computer Science, Millersville University, Millersville, PA 17551 USA
elzer@cs.millersville.edu

Abstract. Information graphics, such as bar charts and line graphs, that appear in popular media generally have a message that they are intended to convey. We have developed a novel plan inference system that uses evidence in the form of communicative signals from the graphic to recognize the graphic designer's intended message. We contend that plan inference research would benefit from examining how each of its evidence sources impacts the system's success. This paper presents such an evidence analysis for the communicative signals that are captured in our plan inference system, and the paper shows how the results of this evidence analysis are informing our research on plan recognition and application systems.

1 Introduction

Plan recognition systems develop a model of an agent's plans and goals by analyzing the agent's actions. We contend that plan recognition research and its applications would be strengthened by focusing not only on the success of the overall system but also on the impact of the different evidence sources on the system's ability to form a correct hypothesis. This paper describes a novel use of plan recognition — namely, to hypothesize the intended message of an information graphic. The paper presents an analysis of the impact of different communicative signals on the system's success, and it discusses how our research has benefited from this evidence analysis.

Section 2 introduces plan recognition from information graphics. Section 3 presents our Bayesian model of plan recognition, with emphasis on the cues available to a graphic designer. Section 4 presents an analysis of the various types of cues on the system's recognition of a graphic's message; Section 5 discusses the impact of this evidence analysis on our work and argues that other plan recognition research would benefit from evaluating the contributions of their various evidence sources.

[⋆] This material is based upon work supported by the National Science Foundation under Grant No. IIS-0534948

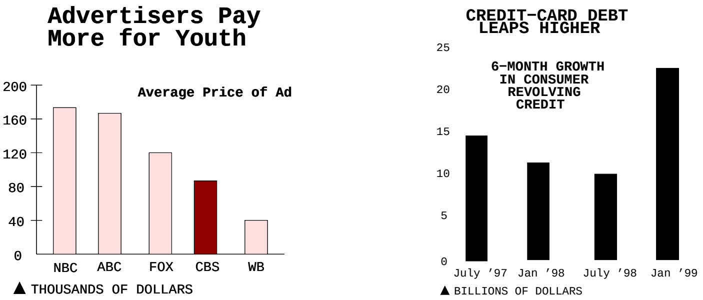


Fig. 1. Two Graphics from Business Week

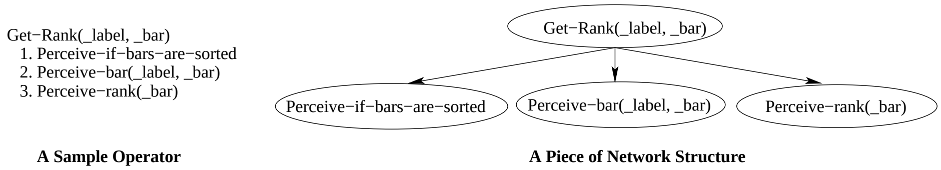
2 Plan Inference and Information Graphics

Our research is concerned with information graphics (non-pictorial graphs such as bar charts and line graphs). Most information graphics that appear in popular media such as magazines, newspapers, and formal reports, have a message that they are intended to convey. Consider for example the information graphics displayed in Figure 1. The intended message of the left graphic is ostensibly that CBS ranks fourth in terms of the average price of Ad compared with NBC, ABC, FOX, and WB, and the intended message of the right graphic is ostensibly that consumer revolving credit grew in Jan '99 in contrast with the previously decreasing trend from July '97 to July '98.

We have developed a novel application of plan inference techniques to information graphics. In the context of our work, the designer of the graphic is treated as the user whose plan is being modeled, and plan inference hypothesizes this plan that the graphic designer intends for the viewer of the graphic to infer in recognizing the intended message of the graphic. This correlates with plan inference in language understanding, where the speaker intends for the listener to infer the speaker's plan and thereby recognize the intended meaning of the speaker's utterance. And as with language understanding, identifying the intended message of an information graphic will enable our system to exhibit behavior appropriate to the recognized message.

3 Bayesian Plan Recognition from Information Graphics

We have designed a Bayesian system for inferring the plan that the graphic designer intends for the viewer to pursue in recognizing the graphic's message which is captured by the plan's top-level communicative goal. Although we believe that our methodology is extendible to other kinds of information graphics, our implemented system currently handles only simple bar charts such as the ones shown in Figure 1. Input to our plan inference system is an xml representation of a graphic, produced by a computer vision module[1] that specifies the graph's axes, the individual bars (including their heights, labels, color, etc.), and the graph's caption. The plan inference system outputs a logical representation of the intended message of the graphic which is then realized in English.



```
Get−Rank(_label, _bar)
  1. Perceive−if−bars−are−sorted
  2. Perceive−bar(_label, _bar)
  3. Perceive−rank(_bar)
```

A Sample Operator

A Piece of Network Structure

Fig. 2. A Sample Operator and its Associated Piece of Network Structure

3.1 Constructing the Network

The top level of our Bayesian network captures the twelve categories of communicative goals (or categories of messages) that we identified for simple bar charts, such as *getting the rank of an entity*, *comparing two entities*, *contrasting a point with a trend*, etc. As with previous plan recognition work[2], we use operators to decompose high-level goals into a set of subgoals; since we are working with information graphics, subgoals eventually decompose into perceptual or cognitive tasks[3], where a perceptual task is one that can be performed by viewing the graphic (such as determining which of two bars is taller in a bar chart) and a cognitive task is one that requires a mental computation (such as interpolating between two values). The operators determine the structure of the Bayesian network, in that the subgoals in an operator become children of their goal node in the Bayesian network. Figure 2 displays a plan operator for getting the rank of a bar given its label and the piece of network structure derived from it. However, memory limitations restrict the size of the network. Our solution is to start with only the ten easiest perceptual tasks as identified by our effort estimation rules[4] (limited to one instantiation per task type) and with perceptual tasks whose parameters are salient entities (such as a bar that is colored differently from other bars, as in Figure 1). The network is then built by both 1) chaining backwards from these primitive perceptual tasks to higher-level goals, and 2) chaining forwards from each newly entered node to primitive tasks.

3.2 Evidence Nodes

Bayesian networks need evidence for guiding the construction of a hypothesis. We have identified eight kinds of communicative signals that can appear in information graphics: effort, highlighting, annotation, most-recent-date, salient-height, noun-matching-bar-label, verb, and adjective.

The AutoBrief project was concerned with generating information graphics[3]. We have adopted their hypothesis that the graphic designer constructs a graphic that makes intended tasks as easy as possible. Thus the relative difficulty of different perceptual tasks serves as a communicative signal about which tasks the viewer was intended to perform in deciphering the graphic's intended message. For example, identifying the taller of two bars in a bar chart will be much easier if the bars are adjacent and significantly different in height than if they are widely separated and only slightly different in height. We constructed a set of effort estimation rules for estimating the effort involved in performing different perceptual tasks on simple bar charts. These rules have been validated by eyetracking experiments and are presented in [4].

Coloring one bar differently from other bars in the bar chart, or annotating it with a special mark, draws attention to the bar and provides highlighting or annotation evidence. The presence of a bar associated most closely (via its label) with the date of the publication is used as most-recent-date evidence, since we hypothesize that it is mutually believed that the viewer will notice events that are current. A bar that is significantly taller than other bars "stands out" in the graphic, and provides salient-height evidence. The presence of a noun in the caption that matches the label of a bar in the graphic is a communicative signal that the referenced entity is important to the graphic designer's message.

Nodes capturing these six types of evidence are attached to each primitive perceptual task in the network, since effort evidence captures the difficulty of a perceptual task and the other five kinds of evidence capture the presence/absence of some feature of a bar serving as a parameter of the perceptual task.

The presence of certain verbs (such as *lag* or *rise*) and adjectives (such as *more* or *largest*) in the caption can signal the category of the intended message, such as conveying the rank of an entity or conveying a rising trend. (Adjectives derived from verbs, such as *rising*, are treated as verbs.) We use a part-of-speech tagger and a stemmer to identify the presence of one of our identified verb or adjective classes in the caption; nodes capturing this evidence are attached to the top-level node in the network since they suggest a general category of message.

3.3 Implementation

The conditional probability tables in our Bayesian network are obtained from our corpus of 110 bar charts. To facilitate leave-one-out cross validation of results (and also re-training under different sets of evidence, as discussed in the next section), we automated the construction of a spreadsheet containing the information needed from each graphic to compute the necessary probabilities. System performance was measured using leave-one-out cross validation. The system's hypothesis for a graphic was viewed as correct if it matched the intended message assigned to the graphic by the human annotators and the probability that the system assigned to the hypothesis exceeded 50%. Overall success was computed as the average success over the 110 graphics in the corpus.

4 Analyzing How Evidence Impacts Plan Recognition

Research in many areas, including dialogue act tagging[6], emotion recognition[7], and question answering[8], have analyzed their knowledge sources to identify to what extent each affects the system's hypothesis. In many cases, this has consisted of examining the features in the resulting decision tree or comparing performance results of decision trees constructed from different sets of features; in the work on question answering by Moldovan et. al., the system is prevented from accessing various resources such as WordNet, and system performance is compared to a baseline system with all resources accessible. However, in the domain of plan recognition, evaluation has focused on the overall success of the system and has given little attention to how much each evidence source contributes to recognizing the user's plans and goals. We contend that an analysis of the impact of the various sources of evidence can inform subsequent research directed at improving the system and can be used in the development of applications utilizing plan inference. This section provides an analysis of the contribution of each of our evidence sources to recognizing the intended message of an information graphic, and Section 5 discusses how this analysis has impacted our subsequent research.

We wanted to evaluate how each kind of evidence impacted system performance by 1) examining system performance with only one kind of evidence, and 2) examining the degradation in system performance when a particular kind of evidence is disabled. It is important to note that disabling an evidence source means that we effectively remove this kind of evidence node from the network by eliminating its ability to contribute to the network probabilities. This is different from recording that the particular cue, such as highlighting, is absent, since the absence, as well as the presence, of a cue is evidence.

To provide baselines for our experiments, we ran the system first without any evidence sources enabled and then with all eight evidence sources enabled. Even without any evidence sources, the system still has certain basic information, such as the ten easiest perceptual tasks (limited to one instantiation per task type) from which the Bayesian net is constructed and whether the independent axis is ordinal (such as consecutive dates, age groups, etc.). The system without any evidence sources enabled had a success rate of only 6% at identifying the intended message of a bar chart, while the system with all evidence sources enabled had a success rate of 79%.

We then ran eight experiments in which only one kind of evidence (such as the presence/absence of highlighting in a graphic) was enabled, and compared the improvement in performance with the baseline system with no evidence sources enabled. Similarly, we ran eight experiments in which one kind of evidence was disabled, and analyzed the degradation in performance (if any) that resulted from omission of this evidence source. We used a one-tailed McNemar test for the significance of changes in related samples[9, 5]. McNemar is a non-parametric test that is appropriate when the samples are related. For our experiments, the samples are related since one sample is obtained from a baseline system and the other sample is obtained after some perturbation of the system (by adding or removing an evidence source). The results of these experiments are shown in Tables 1 and 2. In Table 1, the hypothesis H_1 is that adding the particular evidence source produces better performance than the system with no evidence; in Table 2, H_1 is that removing an evidence source results in worse performance. The rightmost column of each table gives the p value — that is, the significance level at which the null hypothesis is rejected and H_1 is accepted.

Table 1. Improvement in Performance with Addition of Evidence Source

Baseline: System Without Any Evidence	6% success rate		
TYPE OF EVIDENCE ADDED	**SUCCESS RATE**	**McNEMAR STATISTIC**	**p VALUE**
Only effort evidence	57%	52.155	.0001
Only current-date evidence	49%	45.021	.0001
Only annotation evidence	35%	29.032	.0001
Only verb evidence	24%	17.053	.0001
Only highlighting evidence	21%	14.063	.0001
Only evidence about salient-height	19%	12.071	.0005
Only evidence about noun-matching-bar-label	18%	9.600	.001
Adjective	14%	6.125	.01

Table 1 shows that addition of every evidence source produces improved performance that is statistically significant at the .01 level or better. The three evidence sources producing the largest improvement in performance were *effort*, *current-date*, and *annotation*. On the other hand, Table 2 shows that *noun-matching-bar-label*, *effort*, and *current-date* are the only evidence sources whose removal caused degradation in performance that was statistically significant at the .01 level or better.[3] Moreover, the degradation in performance was much less than the contribution of each of these evidence sources when they are the only source used. Thus it is clear that the evidence sources compensate for one another: when one source of evidence is disabled, cues from other sources generally provide evidence that still enables recognition of the intended message.

We will discuss the *effort* and *noun-matching-bar-label* evidence sources since their removal has the greatest impact on system performance. *Effort* both has the greatest impact on system performance when it is the only source of evidence and results in major degradation in performance when it is removed. Although we did not expect this, in retrospect it is not surprising since *effort* evidence reflects how the organization of data in the graphic facilitates different perceptual tasks; thus it affects the message of every graphic whereas other signals, such as *highlighting*, only occur in some graphs. However, *effort* by itself is insufficient for recognizing some kinds of messages, such as that a graph is conveying the rank of a particular bar. (The rules for estimating *effort* do not take salience into account; thus a bar being highlighted does not affect the *effort* computation, but the highlighting is captured by the *highlighting* evidence node.) We also find that, when *effort* is the only evidence source, the average probability attached to the correct hypotheses is 70% whereas the average probability assigned to hypotheses about these same graphs with all evidence is 98%. Thus we conclude that although *effort* has a strong impact on system performance, not only is it insufficient by itself for recognizing certain categories of intention but it results in less confidence assigned to the correct hypotheses that it does produce.

[3] Note that disabling the *adjective* evidence source improved performance, although this change was not statistically significant.

Table 2. Degradation in Performance with Omission of Evidence Source

Baseline: System With All Evidence	79% success rate

	SUCCESS	**McNEMAR**	**p**
TYPE OF EVIDENCE OMITTED	**RATE**	**STATISTIC**	**VALUE**
Noun-matching-bar-label evidence	70%	8.100	.005
Effort evidence	71%	5.818	.01
Current-date evidence	72%	6.125[4]	.01
Highlighting evidence	74%	3.125	.05
Salient-height evidence	74%	3.125	.05
Annotation evidence	75%	2.250	∗
Verb evidence	78%	0.500	∗
Adjective evidence	81%	0.500	∗

∗ Not statistically significant

Noun-matching-bar-label is another evidence source whose omission results in large degradation in system performance. We examined the graphs whose captions contained a noun matching a bar label and whose intended message was correctly identified using all evidence. Without *noun-matching-bar-label* evidence, the system failed to identify the correct message when there was no other evidence that made the bar salient, such as highlighting of the bar or the bar being significantly taller than other bars. However, in ten graphs, such additional evidence enabled the system to recognize the intended message even when *noun-matching bar-label* evidence was disabled. Thus we see that the absence of *noun-matching-bar-label* evidence degrades system performance, but this degradation is sometimes alleviated by the presence of other compensating evidence.

5 Lessons Learned

We contend that research on plan recognition and its use in adaptive systems would benefit from examining the impact of the individual evidence sources on the system's performance. In this section, we support this contention by showing how our evidence analysis has informed our research.

5.1 Applications of Plan Recognition from Information Graphics

We are applying plan inference from information graphics to several projects. In the area of digital libraries, the graphic's intended message will be used as the basis for the graphic's summarization, indexing, and retrieval; furthermore,

[4] The McNemar statistic is based on 1) the number correct by System-1 and wrong by System-2, and 2) the number wrong by System-1 and correct by System-2. Thus although a greater difference in success rates usually correlates with greater statistical significance, this is not always the case.

the graphic's summary will be integrated into an overall summary of the multimodal document. In the area of assistive technology, we have built a system, SIGHT, that infers the graphic's intended message and conveys it via speech to individuals with sight-impairments. A third project is a graph design assistant that will compare the message inferred for a graphic with the designer's intentions and help the designer improve the graphic so that it better conveys his desired message. And lastly, we are investigating a system for tutoring individuals with disabilities in the analysis, understanding, and construction of information graphics.

5.2 Implications of Evidence Analysis for Plan Recognition

Recognizing a graphic's intended message is an integral part of each of our projects; consequently, improving our system's success at plan recognition and extending our methodology to more complex graphics, such as grouped bar charts, is important. *Effort* evidence requires the construction of effort estimation rules and their validation via eyetracking experiments with human subjects; thus it requires substantial research, particularly in the case of grouped bar charts since there is little prior work by cognitive psychologists to draw on. Contrary to our expectations prior to our evidence analysis, *effort* evidence has the strongest overall impact on system performance, (in terms of its contribution when it is the only evidence source and the degradation in system performance when *effort* evidence is disabled). Thus our evidence analysis has caused us to give high priority to devising very good effort estimates for complex graphics.

Disabling *noun-matching-bar-label* evidence also had a major impact on system performance. This suggested that we examine our graphics to determine whether any similar forms of evidence were overlooked in our implementation. We found that mutual beliefs by the graphic designer and the intended viewer about implicitly salient entities seems to play a role in the intended message of a graphic. These implicitly salient entities are a function of the intended audience of a publication. For example, *Canadian Business* is directed toward Canadians. Thus, implicitly salient entities are those associated with Canada, such as *Canada*, *Toronto*, any Canadian company, etc. We hypothesize that if only one bar in a bar chart is labelled with an implicitly salient entity, this salience is similar to mentioning the bar's label in the caption. This conjecture is supported by an analysis of the accompanying articles of such graphics, where it is clear from the article that the graphic designer intended that the implicitly salient entity play a major role in the graphic's message. Thus we are adding such implicitly salient entities as a new evidence source.

We expected *verb* evidence to be a major factor in system success, and had begun to study WordNet similarity metrics that might improve system performance by identifying when new verbs in captions were related to our identified verb classes. However, our evidence analysis (particularly Table 2) suggests that additional verb evidence will not have much of an impact on system performance. Upon reviewing our graphics, we found that there is too much contradictory evidence provided by verbs; for example, the caption on a recent graphic conveying a rising trend in revenue from water parks was entitled *Slip Slidin' Away* — the verb *slide* would be most associated with falling trends and thus hamper recognition of the graphic's intended message. Thus our evidence analysis has led us to conclude that additional work on verb evidence would not be a productive use of research resources.

Our evidence analysis also motivated an addition to our system's message categories. When we gave our system a new bar chart containing a large number of bars and with the bar for Canada highlighted, it failed to infer that the graph was conveying the rank of *Canada*. Since our evidence analysis indicated that *effort* evidence has the strongest impact on system performance, we looked at the effort estimates for the perceptual tasks involved in the plan for the Get-rank message, and we found that identifying the exact rank (14th) of *Canada* required considerable effort given the large number of bars. Upon further reflection and discussion with viewers of the graphic, we realized that the graphic was not conveying the *exact* rank of Canada, but rather its *relative* rank (low, middle, high); estimating *relative rank* is a much easier perceptual task than computing *exact rank*. Thus we are adding *Get-relative-rank* as a new message category.

5.3 Exploiting Evidence Analysis in Applications

In addition to influencing plan inference research, evidence analysis can guide application projects by suggesting which sources of evidence will be most useful. Our graph design assistant will use the results of the evidence analysis to suggest ways in which a graphic might be improved so that it better conveys the designer's intended message. Evidence that has the strongest impact on plan inference, both overall (such as *effort* evidence) and with respect to the specific desired message category, will be considered first in deciding how the graph might be improved.

Our graph retrieval system for digital libraries will respond to requests for a particular kind of graphic. If the library does not contain a graphic whose intended message matches the request, we anticipate using the relative contribution of the different evidence sources to rank other graphics from which the desired information can be inferred. For example, suppose that the system is unable to satisfy a request for a graphic whose intended message is the rank of the CBS network in terms of revenue, but the system does have two alternative graphs from which the desired information could be inferred: 1) a graph conveying the rank of the NBC network (with the bar for NBC highlighted and the bar for CBS not distinguished in any way), and 2) a graph with the bars for network revenue ordered alphabetically by network rather than ordered by bar height. Since highlighting a bar has less impact on plan inference than does perceptual effort, the first alternative would be ranked higher than the second. Furthermore, the ranking of the different evidence sources will be used to explain why this graphic was selected.

Our SIGHT system provides blind individuals with access to information graphics by conveying the graphic's intended message via speech. The system should be able to justify its inferred message upon request, rather than forcing a blind individual to accept without question what the system has produced. The results of our evidence analysis will affect which evidence sources are considered first in constructing the justification. And lastly, our system for tutoring individuals with learning disabilities will use the results of our evidence analysis to order the kinds of evidence that students are taught to consider in inferring the graphic's message and for teaching students to construct graphs that effectively convey their desired message.

6 Conclusion

This paper presented our implemented system for extending plan recognition techniques to inferring the intended message of one kind of information graphic, simple bar charts. Prior work on plan recognition has focused on the success of the overall system, without considering the impact of different evidence sources. We have analyzed the individual evidence sources in our system, both in terms of their contribution to system performance when they are the only enabled evidence source and in terms of degradation in system performance when they are disabled. We contend that the results of such evidence analysis should be taken into account in further research, and we have shown the impact that our evidence analysis has had (and is having) on our plan inference in the domain of information graphics and on our application projects.

References

1. Chester, D., Elzer, S.: Getting computers to see information graphics so users do not have to. Proc. of 15th Int. Symp. on Methodologies for Intelligent Systems. (2005)
2. Perrault, R., Allen, J.: A Plan-Based Analysis of Indirect Speech Acts. American Journal of Computational Linguistics **6**(3-4) (1980) 167–182
3. Kerpedjiev, S., Roth, S.: Mapping communicative goals into conceptual tasks to generate graphics in discourse. In: Proc. of Intelligent User Interfaces. (2000) 60–67
4. Elzer, S., Green, N., Carberry, S., Hoffman, J.: A model of perceptual task effort for bar charts and its role in recognizing intention. User Modeling and User-Adapted Interaction (2006) 1–30
5. GraphPad Software. QuickCalcs: Online Calculators for Scientists (2002). http://www.graphpad.com/quickcalcs/McNemarEx.cfm
6. Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C.: Can prosody aid the automatic classification of dialog acts in conversational speech? Language and Speech (1998)
7. Forbes-Riley, K., Litman, D.: Predicting emotion in spoken dialogue from multiple knowledge sources. In: Proceedings of the HLT/NAACL. (2004) 201–208
8. Moldovan, D., Pasca, M., Harabagiu, S., Surdeanu, M.: Performance issues and error analysis in an open-domain question answering system. In: Proc. of the 40th Annual Meeting of the Assocation for Computational Linguistics. (2002) 33–40
9. Daniel, W.: Applied Nonparametric Statistics. Houghton Mifflin (1978)
10. Kennel, A.: Audiograf: A diagram-reader for the blind. In: Second Annual ACM Conference on Assistive Technologies. (1996) 51–56