Presentation Schemes for Component Analysis in IR Experiments

Paul Kantor[1] and Jimmy Lin[2]

[1]School of Communication, Information and Library Studies
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901
kantor@scils.rutgers.edu

[2]College of Information Studies
University of Maryland
College Park, MD 20742
jimmylin@umd.edu

1. Introduction

Information retrieval research, at least as conceived by the SIGIR community, is fundamentally experimental in nature. As such, the presentation of results from controlled, reproducible experiments lies at the core of our work. Many reports follow the same general format: authors propose a new retrieval method, whose performance on some well-defined task is compared against a baseline. Authors also report results from alternative configurations, e.g., variations in parameters, turning off (ablation) of different components, etc. The presentation of experimental results forms an integral part of the conferences and journals that comprise the medium in which knowledge is disseminated.

The table is the form most often used to convey results. In its standard use, each row represents a particular system variant and each column represents a particular metric. The rows are often arranged in increasing complexity of the system variants so that the performance metric trends upward as one scans down a column. In addition to absolute values, the columns usually report relative increases with respect to some baseline. Finally, little decorative icons are used to indicate statistical significance.[1] Although we note that there are alternatives for presenting results, for example, bar graphs, and triangular matrices for significance tests (Belkin et al., 1995), the table is by far the most common presentation scheme.

This short article aims to challenge "tabular tyranny", arguing that tables are often not the best presentation scheme for IR experiments. In cases where one must convey complex relationships between individual components, alternative schemes may be preferable. Here, we present some possible alternatives, and hope to stimulate the community into thinking more seriously about such issues.

2. Presenting Data in Tables

For the sake of illustration, we consider data from Lin and Demner-Fushman (2006).[2] The technical content is "not material" here, but a brief overview of the experimental setup helps to highlight our ideas. That paper is about the use of a semantic model for retrieval in the domain of clinical medicine. Table 1 shows the comparison (in terms of MAP) between the semantic model and the Indri system. Absolute metrics, relative differences, and statistical significance are noted (** for $p<0.01$, * for $p<0.05$). The proposed semantic model can be further decomposed into two components: frame

[1] Like these: $^{\dagger\ddagger*\blacktriangle\circ\diamond}$.

[2] In fact, this work arose from discussions about the implications of results presented therein.

structure and task knowledge. Table 2 shows the relative performance contribution of each component. To quantify how much a component contributes to the overall semantic model, as well as comparisons to Indri (the baseline), differences relative to each are displayed in separate columns. Finally, Table 3 shows the effects of "partial" semantic models, or ablation analyses, by combining Indri with either the frame structure or task knowledge component separately. These tables are typical of how IR results are often presented.

	MAP
Indri	0.544
Semantic Model	0.718 (+32.1%)**
Indri + Semantic Model	0.730 (+34.3%)**

Table 1: Comparison between Indri and a semantic model for retrieval in the medical domain.

	MAP	vs. Semantic	vs. Indri
Frame	0.646	–10.0%**	+18.8%*
Task	0.620	–25.1%**	–1.1%n.s.

Table 2: Individual contributions of the two primary components in the semantic model.

	MAP
Indri	0.544
Indri + Frame	0.668 (+22.9%)**
Indri + Task	0.620 (+14.0%)**

Table 3: Combining Indri with components of the semantic model.

One notices an "impedance mismatch" between the conceptual organization of the experiments and the presentation of results. The central question of the work is "how does a retrieval method based on semantic domain knowledge (which can be decomposed into frame structure and task knowledge) compare to existing methods (i.e., Indri)?" Three separate tables are required to compare various components in different combinations. Yet, many interesting questions are not readily answered by them. For example, a tabular display is unable to flag all pairwise differences that are statistically significant, since in most cases comparison is made only to the baseline. The linear display of experimental conditions in rows is insufficient to capture conceptual relations between system combinations (for example "A+B" vs. "A+C"). In addition, magnitudes are difficult to illustrate, thus leading many authors to use devices like bold or italic font to draw attention to certain figures (e.g., highest value).

3. Two Alternatives

We present two alternative presentation schemes that are graphical, rather than textual, in nature. We set out with three goals: (1) to clearly show the amount of improvement attributable to each system component; (2) to show the effect of component interactions. This is challenging since the number of potential interactions grows quickly with the number of distinct components. (3) Furthermore, we seek a method for clearly indicating the statistical significance of any improvements, once again noting that the possible number of comparisons grows as the square of the number of experimental

conditions. (4) To round out the desiderata, we constrain the design to static schemes, suitable for appearance in print. Although allowing interaction would greatly expand the solution space, such alternatives are not practical given the dominant mode of publication today. For the same reason, we avoid excessive use of color, instead favoring schemes that reproduce well in grayscale.

We propose two types of diagrams, shown in Figures 1 and 2, that graphically encode at least the information presented in Tables 1, 2, and 3. We build up the idea in a series of steps.

3.1 The Baseline

First, we consider the role of the baseline. The value "0" is never taken as the baseline performance, which should be based on some point the authors (if not the community) take as the current "state of the art". Ideally, this should be something like "performance achieved by the best open source algorithm on this particular data set".

3.2 Relative Performance

In addition to absolute performance values (e.g., mean average precision), it is customary to report some sort of relative performance. Most often one sees "percentage improvement over baseline", which we do not believe to be sensible in many cases. We propose instead that progress be reported in terms of the advance toward a conceivable performance of 100% (or whatever value represents "perfection").[3]

The benefit of this approach is that it accurately reflects the increasing difficulty of making ever smaller advances as the room for improvement decreases. This stands in contrast to the custom in the IR literature of reporting improvement as a percentage of the current baseline. The effect is that an increase from, say, 10% to 20% would look huge, while the much more important increase from 89% to 99% would seem far smaller. Researchers wish to make improvements look large, which is natural due to the experimental nature of IR. However, John McCarthy's observation about AI is apposite : "…[computer chess] has developed much as genetics might have if the geneticists had concentrated their efforts starting in 1910 on breeding racing *Drosophila*. We would have *some* science, but mainly we would have very fast fruit flies" (McCarthy, 1997). But as we approach the limits of computational methods, it seems more attractive to talk in terms of the "errors avoided", which is precisely what the proposed measure does. Thus, in addition to the absolute performance, we report a metric we dub (with tongue somewhat in cheek) AI (that is, Achievable Improvement): $AI = (p_{new} - p_{base}) / (1 - p_{base})$. Note that this idea is not unrelated to "percent of monolingual performance" often cited in CLIR experiments, although we advocate a broader, consistent application.

3.3 Displaying Component Contributions and Significance of Differences

Next, we address the issue of simultaneously displaying the performance contributions of system components and the statistical significance of the observed differences. This is accomplished by arranging individual components along the x axis and using the y axis to represent performance. When the number of components is small (say, three), arrows can be used directly to denote the combinations (Figure 1). We refer to this as the DAG (Directed Acyclic Graph) variant. When there are more than three components, we favor a grid display that employs an implicit binary notation to denote combinations (Figure 2), since the proliferation of arrows would result in too much clutter. We refer to this second diagram as the BCM (Binary Component Matrix) variant.

[3] This idea is being explored in a dissertation defended by Robert Rittman, where the variant used is called "Accuracy Gain". (Rutgers, April 13, 2007; personal communication).

For the DAG variant, statistical significance can be directly encoded in the arrows. Results of additional significance testing can be conveyed in a series of bars on the right edge of the diagram. The ends of a bar indicate which conditions are being compared. The thickness of the bar indicates the level of significance. As an example, the difference between "Indri" (MAP=0.544) and "frame" (MAP=0.646) is significant at the 95% level. For the BCM variant the bar display alone is preferred.

4. Interpreting the Results

Let us examine Figure 1 in more detail. The heavy baseline represents "Indri", which is taken as the prior state of the art. Note that of the two components, "frame" by itself is substantially better, while "task" alone is slightly worse. The arrows make it clear that all two-way combinations are synergistic. The diagram also shows "negative interaction effects", in that the overall improvement obtained by combining two methods is always less than the sum of individual improvements. Generally, as we move up the chart, arrows become more horizontal (diminishing returns) but the addition of components always provides some improvement (although it is not necessarily statistically significant).

We now turn our attention to Figure 2, the BCM variant. The solid dots mark which components are active in the combination that produces a certain level of performance. Unlike the DAG variant, the BCM diagram is arbitrarily extensible to any number of components, although information about the effect of incrementally adding individual components is more difficult to extract.

An interesting issue concerns the top edge of the graph. For "truth in advertising", we set the upper bound at "perfection", which unfortunately leaves a large white space at the top. Such egregious waste would surely offend Tufte (2001), but it signals clearly that these improvements only solve half of the problem and that we still have a long way to go. Nevertheless, the space can be put to good use showing a block diagram of the system, portraits of the authors (as we have), or some other "eye candy".[4]

5. Conclusion

We wish to close with a few thoughts on potential barriers to adoption. The accumulation of a critical mass is an important issue, but one we believe can be jump-started with a few influential early adopters. Training of readers also poses a challenge; at least initially, we envision papers having rather elaborate explanations of these schemes. Finally, we'll need tools to facilitate the creation of these diagrams.

To sum up: We believe in the importance of presentation schemes that *transparently* convey insights from IR experiments. In many cases, tables are insufficient for this task. We describe two possible solutions; no doubt there are other alternatives as well. In any case, we hope to call attention to presentation of results as a problem worth reconsidering as Information Retrieval enters "middle age".

6. Acknowledgements

We'd like to thank colleagues who commented on an earlier draft of this paper, particularly Doug Oard. Of course, we are solely responsible for all remaining errors and places where the humor may prove heavy-handed. ☺ PK particularly thanks Prof. Oard for a life-transforming lift in his Mooney. As usual, JL thanks Kiri and Esther for their kind support.

[4] For the record, we do not consider portraits of the authors to be "eye-candy". Nevertheless, we feared that the portrait of an attractive model would not be recognizable to the IR community.

7. References

N. J. Belkin, P. Kantor, E. A. Fox and J. A. Shaw. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management*, 31(3):431-448, 1995.

J. Lin and D. Demner-Fushman. The Role of Knowledge in Conceptual Retrieval: A Study in the Domain of Clinical Medicine. *Proceedings of SIGIR 2006*, pp 99-106, 2006.

J. McCarthy. AI as Sport. *Science*, 276(5318):1518-1519, June 1997.

E. Tufte. The Visual Display of Quantitative Information, 2nd edition. Graphics Press: Cheshire, Connecticut, 2001.

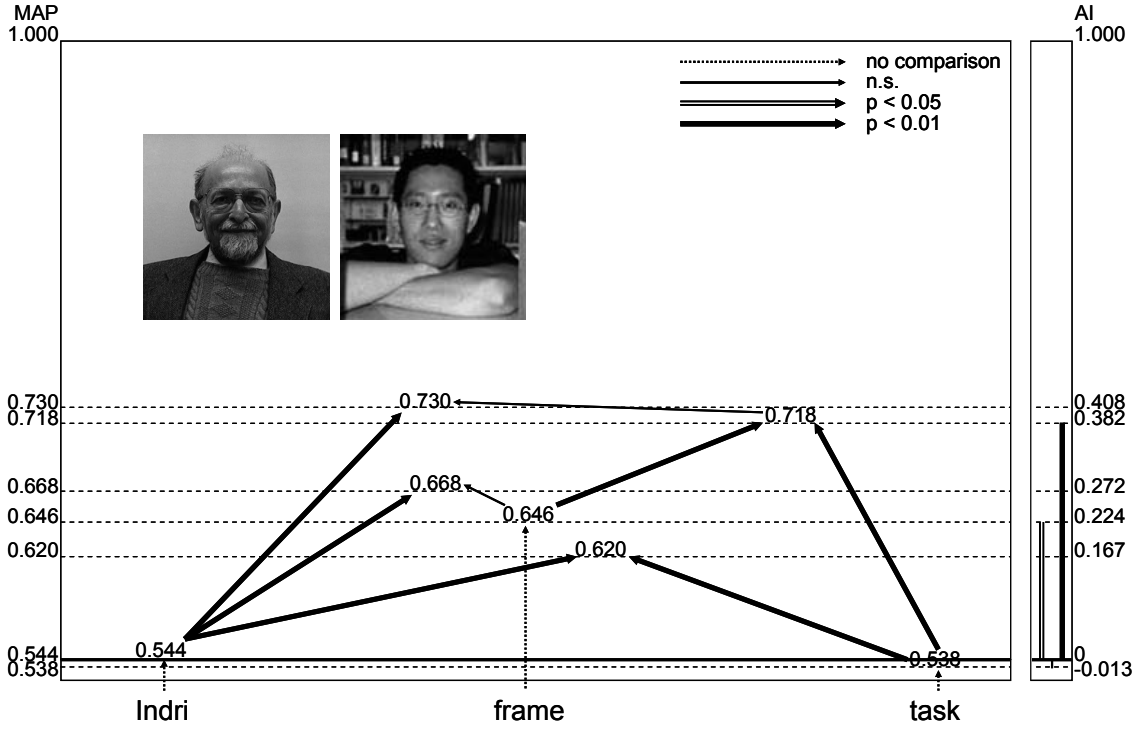

Figure 1: The DAG variant of our presentation scheme. Component combinations are denoted explicitly by arrows, which also flag the significance of improvements. "Frame" and "task" are two components of the semantic model.

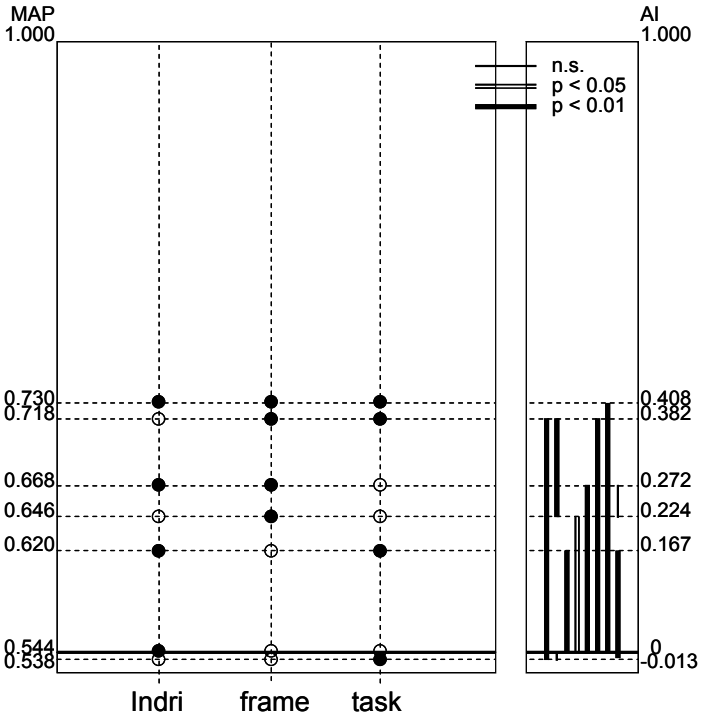

Figure 2: The BCM variant of our presentation scheme, where each horizontal sequence of dots can be understood to denote a particular system combination (e.g., solid = "on", empty = "off").