

Probabilistic and Evolutionary Aspects of Online Bibliographic Retrieval: An Exploratory Investigation

Joseph Janes

School of Information and Library Studies, University of Michigan, Ann Arbor

Presented at SIG/CON, October 19, 1994, Alexandria, VA

Background

This research focuses on the roles that evolution and probability play in information retrieval. There has been surprisingly little research done in this area. We are familiar with the work of Bates, Fidel, Saracevic and Kantor, among others, on the characteristics of good online searches and good online searchers, but there is nothing explicitly on the topic that we were interested in. We were motivated by examining some very old literature and philosophy, and an old but yet untested theory (as far as we know) regarding the role of evolution and probability in information work. We've adapted this topic to this particular interest and domain and we are pleased to report on our study of online searching by an infinite number of chimpanzees.

We had difficulty receiving funding for this research. We, along with practically everybody else, submitted a proposal to the NSF/DLP, the Digital Lemur Project. We were unsuccessful in that attempt, largely because the Digital Lemur Project is differently focused in many ways. They are using chimpanzees as *relevance judges*, as opposed to online searchers, which is what we were interested in. Preliminary results from the Digital Lemur Project are in and it seems that they are comparable to TREC results.

Given our difficulty in finding funding to pursue this line of work, we decided to use the money we made last year in the study of relevance judgements and identical twins separated at birth (Janes, 1993).

Research Questions

We asked two primary research questions. First of all, is there any evolutionary advantage in online research? Our null hypothesis is that there is not. The mean performance of chimpanzees will be roughly equal

to that of humans. Our alternative hypothesis is that there is some difference. We elected to do a two-tailed test instead of a one-tailed test, as we have no real theoretical basis, especially given the paucity of literature.

Our second research question is "What is the likelihood that at least one of an infinite number of chimpanzees will produce meaningful results in online searches?"

Literature Review

There has been very little work in the role of evolution in the information and library work, but there is a surprising thread over the years. There was a seminal study in 1935 of the use of orangutans as reference desk assistants. In 1967 we find the work on gorillas in technical services, an interesting piece. In the field of computer science we find chimpanzees as systems analysts, in 1972. The 80s saw the rise of information brokering, and some research on rhesus monkeys as freelance information brokers. There is also one article we are aware of, in press at the moment, in *Information Processing & Management*, on baboons as information science faculty.

Methodology

The searchers we used were in fact many thousands of chimpanzees. Because of the financial difficulties, we had to reduce our sample size from an infinite number to only many thousands. We realize that we have lost some power in our statistical testing, but there was nothing to be done. In fact, we really do not know how many we had. They are really much harder to count than one would think. They move around a great deal, chase each

other, chase us, and we chase them. In any event, it was a lot of chimpanzees. It was a real lot. As far as we could tell, there were no twins, and so there were no biases from the binary separation study.

The control group was an online searching class in the School of Information and Library Studies at the University of Michigan. There were 35 students in the class and to even the playing field, we force fed them bananas for a week, to eliminate rival explanations. (They were very distressed, and we believe that further study on force feeding library students bananas might render fruitful findings.).

All the subjects searched using a variety of different tools—Dialog, LEXIS/NEXIS, et al. I should like to thank, for their cooperation in this study, Knight Ridder, Mead Data Central (before the Reed Elsevier acquisition), and Murray's House of Primates in Ann Arbor. We did not have them search using the gopher protocol, although one of the chimpanzees did actually obtain a gopher. The gopher got a good hit in LEXIS, so we kept that.

We had the groups search several topics—a few from TREC, a few from the twin study, and a few constructed for the study. Each subject was given a password and set loose, which in hindsight was a big mistake, as it took several days to retrieve them all.

Data Analysis

We are still analyzing the data, but we did get some interesting preliminary results. Many of the queries entered by chimpanzees retrieved a high number of hits, and occasionally a real word emerged in the search string, which was encouraging. We also saw the occasional nonsense word, a word that really ought to be a word in English, but that does not bear any content.

We did find in analyzing the data, that there was a graduate student in the chimpanzee group, by mistake (we think—it is hard to tell). We had excellent results from a chimpanzee who is now working at the graduate reference desk at the University of Michigan Library.

In general, the humans did out-perform the chimpanzees, which is reassuring, but not by much, which is somewhat unsettling. The humans consistently out-performed the chimpanzees on recall and precision measures, although the chimpanzees did remarkably well in precision. The high standard deviation indicates a great deal of variation in there, but there is obviously something at work.

What we were most struck by is that the chimpanzees seemed to do better on the Internet, especially WAIS, than the humans. The best explanation we have for that is that it really is a jungle out there.

There does seem to be some evolutionary advantage for the chimpanzees. We did get one nearly complete script from *The Comedy of Errors* from a chimpanzee. We thought we might also get *Measure for Measure*, but we did not.

Conclusions

There are some promising results. There does seem to be an evolutionary advantage in online searching—the humans did do better, but not overwhelmingly. The chimpanzees did get good hits, and they appeared to enjoy searching. Several signed up for the online class this term.

This research does suggest some other possibilities for broadening the opportunities for information and library work, especially for our colleagues perhaps at the lower end of the food chain. There are some obvious possibilities here which are worth examining, especially for the canine family. For example, we might suggest the use of boxers as archivists, and setters as administrators (it occurred to us that pointers could be used here as well). Vendors might choose to explore the possibility of using hounds. Inevitably, the species that we think would be most appropriate to use as online searches would, of course, be retrievers.

References

- Janes, Joseph. 1993. *Early binary separation: Implications for information retrieval*. Unpublished paper presented at SIG/CON 1993.