Chapter 4.

RESULTS

The results discussed in this chapter are based on the 190 searches finally obtained.

The analysis was essentially directed at the comparison question -- i.e. how do Weighted and Boolean searches compare. Essentially each variable included was subjected to a basic distributional analysis, and also cross-tabulated with the "Boolean or Weighted" variable. In the case of (nominal or ordinal) category variables, a chi-square test was then applied; in the case of quantitative variables, the Mann-Whitney U test was used.

Because of the time available, not all variables were included in the analysis, and even for those included, not all possibly interesting analyses were performed. There remains, therefore, a quantitiy of data for future researchers to work on. In particular, there is some scope for diagnostic work on the logs, with a view to understanding and categorising the particular situations in which Weighted searching does or does not work well.

Full tables of the results are given in Appendix A8. Tables are numbered to correspond to each of the sub-sections below (e.g. 4.2.1), and the existence of a table in A8 is indicated in the appropriate sub-section. "Significance" in the tables and "p" in the discussion below mean the same thing; a significance criterion of 5% has been used, i.e. a p value of less than 5% has been taken as significant. The following is a brief discussion of each variable analysed. The results are further discussed by category of variable in section 4.6.

4.1. Results from the presearch form

These results are not expected to show any difference between Boolean and Weighted, and serve as background information only.

Among the categories of users (Consultant/ Doctor/ Lecturer/ Post-graduate/ Researcher/ Other) approximately 40% were postgraduate students, with doctor, researcher, other each accounting for a little over 15%. Almost all (85%) of searches were for research purposes. 60% of users described their search topics as Precise; also 60% wanted a broad search. 60% again had previously had online searches done for them; 15% had done searches themselves (without an intermediary).

4.2. Results from the post-search user questionnaire

4.2.1. Indicate your SATISFACTION with the search on the basis of the scale below (excellent/ good/ satisfactory/ poor/ bad).

(Table in A8.)

For both Boolean and weighted, the modal response was Good, with very few Poor or Bad. No significant difference between Boolean and Weighted, although there may be some advantage to Boolean.

4.2.2. Please provide a general assessment of the SEARCH (easy/ average/ difficult).

(Table in A8.)

Mode is Easy; there is absolutely no difference between Boolean and Weighted.

4.2.3. Generally speaking, were the RESULTS of the search: excellent/ good/ satisfactory/ poor/ bad?

(Table in A8.)

For both Boolean and Weighted, the modal response was Good, with very few Poor or Bad. No significant difference.

4.2.4. Please assess the SEARCHER'S CONTRIBUTION to the search (essential/ helpful/ satisfactory/ poor/ bad).

(Table in A8.)

About 75% said Essential. Difference between Boolean and Weighted not significant (p=13%), but a slight tendency to give a lower rating for Weighted.

4.2.5. How close was the online search to your original or intended enquiry (exact/ fairly close/ considerably altered)?

(Table in A8.)

Mode is close; difference not significant (p=12%), but there is some indication of a greater spread (i.e. more extreme values) for Boolean.

4.2.6. Did you GET the number of REFERENCES EXPECTED (less/ as expected/ more)?

(Table in A8.)

Mode is As Expected; no significant difference.

4.2.7. (Weighted searches only) Did you mark any references as relevant?

(No table.)

Of the users who answered this question, almost all said Yes, and about 75% said that it appeared to make the search more effective. However, only about half the users answered the question.

4.2.8. (Boolean searches only) Did you view any references while online?

(No table.)

Similarly, only about half answered the question; of these, almost all said Yes; about 65% said that the search was modified as a result.

4.3. Results from the post-search intermediary's questionnaire

Apart from the questions discussed below, the remaining questions were not expected to show differences between Boolean and Weighted. The average presearch time reported was 18 minutes.

4.3.1. Indicate your OVERALL SATISFACTION with the search (excellent/ good/ satisfactory/ poor/ bad).

(Table in A8.)

The mode is Good. The difference between Boolean and Weighted is not quite significant (p=6%); although each has the same proportion of

Excellent and Good, Weighted gets somewhat fewer Satisfactory and more Poor and Bad ratings.

4.3.2. Please provide a general assessment of the SEARCH PROCESS (easy/ average/ difficult).

(Table in A8.)

Mode is Average; difference not significant (p=17%), but there is apparently a slight advantage to Weighted.

4.3.3. Generally speaking were the RESULTS of the search: excellent/ good/ satisfactory/ poor/ bad?

(Table in A8.)

Again the mode is Good; no significant difference.

4.3.4. What was your REASON FOR FINISHING the search? (Found what was required/ Technical difficulties/ Search strategy failed/ Other)

(Table in A8.)

Virtually all searches indicated "Found what was required", but technical difficulties were also reported in 7% of searches. (It should be mentioned that searches which were aborted at an early stage or before any output was obtained were either re-run or excluded from the sample).

4.4. Results from logs

60% of searches were on Medline only, and 25% on Inspec only; the remainder involved Psychological Abstracts and/or more than one database.

4.4.1. PSS packets sent and received

(Table in A8.)

Skew distribution with a long tail to the right; highly significant difference, with the mean for Weighted more than twice that for Boolean.

4.4.2. Online time

(Table in A8.)

Average 22 minutes. Skew distribution; difference not significant (p=20%), but some indication of a longer time for Weighted: mean for Weighted about 25% higher than for Boolean.

4.4.3. Online citations

(Table in A8.)

Skew distribution; significant difference (p<5%), mean for Weighted 35% less than for Boolean.

4.4.4. Offline citations

(Table in A8.)

Skew distribution; significant difference (p<5%) with fewer for Weighted (mean 30% less than for Boolean).

4.4.5. Terms used in the search, terms added or ammended

(Table in A8.)

Both variables show a significant difference (p<1%), with fewer for Weighted.

4.5. Results from the relevance assessments

4.5.1. Total number of documents assessed

(No table.)

Difference not significant. Note that this number is generally the same as the number of offline citations, except that it is truncated to 50. Average is 36.

4.5.2. Total relevant retrieved

(Table in A8.)

Difference not significant, either for relevance1 (p=11%) or for relevance2 (p=15%). Nevertheless, for both the total is smaller for Weighted (by about 12%).

4.5.3. Precision

(Table in A8.)

Difference again not significant, for either precision1 (p=36%) or precision2 (p=22%). In both cases mean precision for Weighted is slightly higher than for Boolean: 50% against 45% for precision1, 73% against 69% for precision2.

4.6. Summary of results by category of variable

4.6.1. Retrieval effectiveness

Essentially, there appears to be little difference between the systems. There could be a tendency for Weighted searches to achieve higher precision with lower recall, but the difference is within the bounds of sampling error. If there is such an effect, it is likely to relate to the number of terms used (see 4.6.7 below).

The matched-pair data and the novelty data have not yet been analysed.

4.6.2. User effort

The users' subjective assessments of the difficulty of search showed absolutely no difference between weighted and Boolean. Users spent slightly more time on Weighted searching simply because the online time was greater (see next paragraph). Fewer terms were used (see 4.6.7 below) and fewer citations looked at online; on the other hand, the system required an online relevant/non-relevant response in respect of each item looked at.

4.6.3. Cost

Online time was apparently slightly greater for Weighted (though not significantly so), almost certainly because of the large number of Boolean search statements generated by the weighted searching algorithm. This would have affected four separate categories of cost: host charges, telecommunications charges, intermediary's time, user's time. On- and offline citations were fewer with Weighted. It is not quite clear why this should be (though for offline citations, it is consistent with the lower recall result).

4.6.4. Subjective user reactions

There were no significant differences in this category at all. The very slight differences observed do not appear to form part of a consistent pattern.

4.6.5. User characteristics, request characteristics

Previously summarised from the presearch form. Obviously we do not look for differences between the systems in these variables. It might be appropriate to use them as independent variables against which to evaluate changes in performance etc., but this has not been attempted in the present analysis.

4.6.6. Intermediary's contribution

One might look for a reduced intermediary's contribution for Weighted, in that it should (in principle) involve less effort in constructing search statements. In the users' opinion, this may be the case, though the difference observed is not significant. But various indirect pointers suggest something a little more complex. Intermediaries seemed to take the task of learning how to use the system (both at the low level of learning the command language and at the higher level of learning how to make best use of the system) as seriously as any other online search system. They seemed to find it easier for some queries and harder for others; they would perhaps have welcomed weighting techniques as an addition to their range of possiblities, rather than as an alternative to Boolean searching.

4.6.7. Perceived online time

The most consistent group of observed differences between the systems apparently related to the perception of online time on the part of the intermediaries. It was apparent from comments that they felt that weighted searches were taking too long, in particular at the Search command (when the algorithm is generating and transmitting numbers of Boolean statements). Since this time is critically dependent on the number of terms, they tended to cut down the number of terms. This was observed directly, and also probably led to the apparent reduced-recall, increased-precision result for Weighted searches.

It is interesting to note that the average time for Weighted searches was in fact very little higher than that for Boolean. It is suggested that the intermediaries' perception was perhaps a little distorted, probably because they had nothing to do while the Search

command was running.