A note on measuring overlap

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium

Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

M. Goovaerts

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium

Abstract

In measuring the overlap between two sets A and B (e.g. libraries, databases, …) one is obliged to calculate the overlap $\mathrm{O}(\mathrm{A}\,|\,\mathrm{B})$ of A with respect to B (i.e. the fraction of elements of B that are also in A) and $\mathrm{O}(\mathrm{B}\,|\,\mathrm{A})$ of B with respect to A (i.e. the fraction of elements in A that are also in B). Theoretically this requires two samples.

In this paper we explain that one sample can suffice to determine confidence intervals for both $\mathrm{O}(\mathrm{A}\,|\,\mathrm{B})$ and $\mathrm{O}(\mathrm{B}\,|\,\mathrm{A})$.

The paper closes with the example of measuring the overlap between the secundary sources in mathematics MathSciNet and Zentralblatt MATH and with a remark on the estimation of the Jaccard index.

Keywords: overlap; MathSciNet; Zentralblatt MATH; measurement.

Acknowledgement: The authors thank the referees for valuable comments and suggestions.

1. Introduction

The determination of the overlap between two sets (in general) A and B is very important in informetrics. In the case that A and B are libraries, collaborating in a network, overlap determines the degree of the profit one gets from shared cataloguing, since only one catalographic description is needed for the documents in the overlap. Of course, in general, one can study the overlap between n sets (e.g. libraries) $A_1,...,A_n$ but we do not go into this extension here (only a few documents treat this general case – see e.g. [1], [2], [3], [4] and [5]). In case A and B are bibliographic databases (treating similar topics – e.g. mathematics as is the case for MathSciNet and Zentralblatt Math), the need for buying or using both databases increases with decreasing overlap. Also in interlibrary activities, overlap is very important. In principle, if two libraries have a small overlap then one library can benefit very much of using the collections of the other one, except if the collections are on very different topics. But also a large overlap can be interesting in the case one small library is (almost) a subcollection of a large library on the same topic.

But what does it mean “overlap between two sets A and B” ? It is clear that the simple calculation of $\left|A\,\text{Ç}\,B\right|$, i.e. the absolute number of documents in the intersection of A and B, is not very informative since it is not related or normalized to the sizes $\left|A\right|$ and $\left|B\right|$ of A and B, respectively (see Fig. 1). Therefore the following measures of so-called relative overlap are defined (cf. [2], [4], [6], [7] and [8] – for more on overlap we also refer the reader to [9], [10] and [11]): the overlap of A with respect to B:

$$O(A \mid B) = \frac{\left|A\,\text{Ç}\,B\right|}{\left|B\right|} \tag{1}$$

and the overlap of B with respect to A:

$$O(B \mid A) = \frac{\left|A\,\text{Ç}\,B\right|}{\left|A\right|} \tag{2}$$

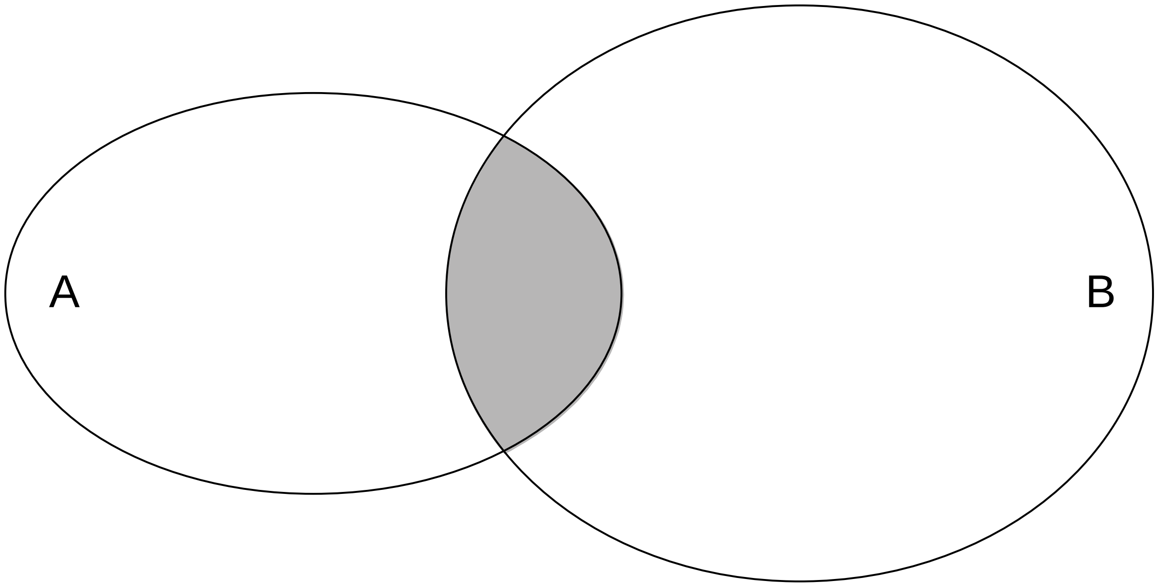

Fig. 1. Absolute overlap between A and B: $A Ç B$.

It is clear that both measures are equally important and that they can be very different e.g. in the case of a small set A being a subset of a large set B – see Fig. 2

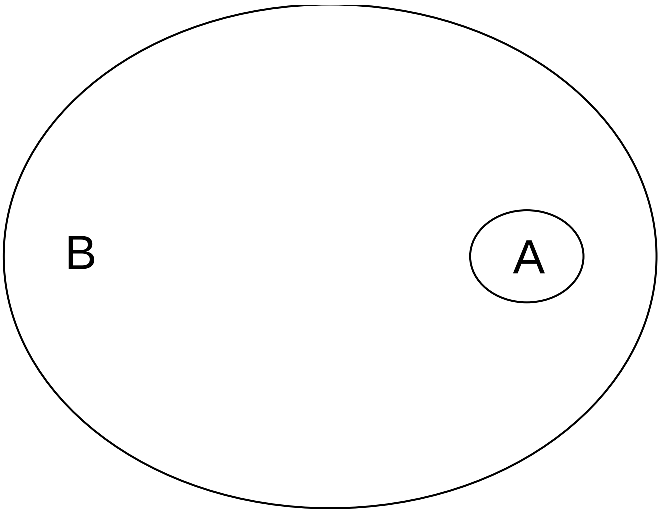

Fig. 2 Large difference between $O(A|B)$ and $O(B|A)$.

Indeed, in this case $O(A|B)$» 0 and $O(B|A) = 1$, the largest possible value. Only in the case that $|A| = |B|$ we have that $O(A|B) = O(B|A)$ as is clear from (1) and (2), but this is not likely to be the case.

How can $O(A|B)$ and $O(B|A)$ be determined ? Of course, if $|A\,\text{Ç}\,B|$, $|A|$ and $|B|$ can be determined, we have $O(A|B)$ and $O(B|A)$ by (1) and (2). Usually $|A|$ and $|B|$ can be determined since library sizes usually are known and since, in the case of a documentary system, information retrieval (IR) techniques can determine $|A|$ and $|B|$ (e.g. by commanding the system to retrieve all documents of a given year and letting the year vary from the initial year until the last year available). The problem is the determination of $|A\,\text{Ç}\,B|$. Even for networked libraries it is virtually impossible to check all documents and for two different documentary systems it is almost impossible as well due to different indexing techniques.

The only realistic way to estimate $O(A|B)$ and $O(B|A)$ is by sampling in A and B and check which documents also belong to the other set. We will explain this method in the next section where we will also determine the confidence intervals for $O(A|B)$ and $O(B|A)$. Next, in the same Section II we will explain how the tedious task of performing two such samples (and looking for the occurrence of the documents of each sample in the other set) can be avoided: we only need one sample and we will also show that it is preferable to take the sample in this set which is the smallest of the two sets (which is also easier due to the possible more homogeneous collection). With this one sample we will still be able to determine two confidence intervals for $O(A|B)$ and $O(B|A)$.

In the third section this methodology is applied to determine the (two-fold) overlap between MathSciNet and Zentralblatt MATH. Also some practical details of this overlap are communicated.

The fourth section presents some conclusions and remarks, e.g. on the estimation of the Jaccard index (called traditional overlap in [12]).

2. The estimation of O(A|B) and O(B|A): two-sample method and improvement to a one-sample method.

2.1. Classical method: two samples.

Let us recall the definition of $O(A|B)$, the overlap of A with respect to B:

$$O(A|B)=\frac{|A\,\text{Ç}\,B|}{|B|} \tag{3}$$

This means: the relative size of $|A\,\text{Ç}\,B|$ with respect to the size of B, hence a fraction of $|B|$. Since we have to determine a fraction of B we have to take a sample (size N_B) in B and determine how many elements in this sample belong to A. We denote the obtained sample fraction by $\bar{x}_{A|B}$. Since a fraction is also an average (indeed: if the sampled document of B is in A we give it a score 1 and if it is not in A we give it a score 0; then the fraction $\bar{x}_{A|B}$ is the average of these scores), we can apply the classical theory of confidence intervals for averages, based on the Central Limit Theorem (cf. [4], [6]) as follows: the unknown population fraction $O(A\,|\,B)$ belongs to the following interval with 95% certainty:

$$O(A\,|\,B) \in \left[\bar{x}_{A|B} - 1.96\sqrt{\frac{\bar{x}_{A|B}\left(1-\bar{x}_{A|B}\right)}{N_B - 1}},\ \bar{x}_{A|B} + 1.96\sqrt{\frac{\bar{x}_{A|B}\left(1-\bar{x}_{A|B}\right)}{N_B - 1}}\right] \tag{4}$$

and similarly (based on the tables of the standard normal distribution) for other confidence intervals (by replacing 1.96 by the corresponding score, e.g. 2.575 for a 99% confidence interval).

This solves the problem of determining $O(A\,|\,B)$ as long as we are able to limit the length of the above interval (4) to satisfactory limits (dependent of the problem), the length of the above interval being

$$L_{A|B} = 2\text{x}1.96\sqrt{\frac{\bar{x}_{A|B}\left(1-\bar{x}_{A|B}\right)}{N_B - 1}} \tag{5}$$

This determines N_B, based on a first, preliminary, sample which determines $\bar{x}_{A|B}$. This is called two-stage sampling but is not the topic of this paper – see e.g. [6].

The estimation of $O(B\,|\,A)$ follows exactly the same lines: now we take a sample in A (size N_A) and determine how many elements in this sample belong to B. We denote the obtained sample fraction by $\bar{x}_{B|A}$. Again we have a confidence interval for $O(B\,|\,A)$: with 95% certainty we have that

$$O(B\,|\,A) \in \left[\bar{x}_{B|A} - 1.96\sqrt{\frac{\bar{x}_{B|A}\left(1-\bar{x}_{B|A}\right)}{N_A - 1}},\ \bar{x}_{B|A} + 1.96\sqrt{\frac{\bar{x}_{B|A}\left(1-\bar{x}_{B|A}\right)}{N_A - 1}}\right] \tag{6}$$

and similarly for the other confidence intervals.

The above method clearly requires two samples and, each time, the (time-consuming) action of checking if these sampled documents belong to the other set. The next subsection is halving this task by deriving a confidence interval for $O(B\,|\,A)$ from the one of $O(A\,|\,B)$ (or vice-versa), by a very simple method.

2.2. Improved method: one sample.

Let us perform the sample to determine $\bar{x}_{A|B}$ and (4) as above, yielding a confidence interval for $O(A\,|\,B)$. Note that

$$O(B\,|\,A) = \frac{|A \cap B|}{|A|}$$

$$= \frac{|A \cap B|}{|B|}\frac{|B|}{|A|}$$

$$O(B\,|\,A) = O(A\,|\,B)\frac{|B|}{|A|} \tag{7}$$

Hence, by (4) and (7) we have that

$$O(B\,|\,A) \in \frac{|B|}{|A|}\left[\overline{x}_{A|B} - 1.96\sqrt{\frac{\overline{x}_{A|B}\left(1-\overline{x}_{A|B}\right)}{N_B - 1}}, \overline{x}_{A|B} + 1.96\sqrt{\frac{\overline{x}_{A|B}\left(1-\overline{x}_{A|B}\right)}{N_B - 1}}\right] \tag{8}$$

Consequently, we obtain the following new confidence interval (in this case 95%):

$$O(B\,|\,A) \in \left[\overline{x}_{A|B}\frac{|B|}{|A|} - 1.96\frac{|B|}{|A|}\sqrt{\frac{\overline{x}_{A|B}\left(1-\overline{x}_{A|B}\right)}{N_B - 1}}, \overline{x}_{A|B}\frac{|B|}{|A|} + 1.96\frac{|B|}{|A|}\sqrt{\frac{\overline{x}_{A|B}\left(1-\overline{x}_{A|B}\right)}{N_B - 1}}\right] \tag{9}$$

and similarly for the other confidence intervals, replacing 1.96 by its corresponding value (e.g. 2.575 for 99% confidence).

Note that this confidence interval only uses the sample to determine $O(A\,|\,B)$. To determine the confidence interval for $O(B\,|\,A)$ it suffices to multiply this interval by $\frac{|B|}{|A|}$. Since the role of A and B is symmetric we can hence advise to choose $|B|$ such that $|B| \leq |A|$ hence such that $\frac{|B|}{|A|} \leq 1$ which guarantees the second interval (9) to be smaller (in absolute sense) than the interval (4). The relative length (with respect to the value $O(B\,|\,A)$) remains the same since also the value $\overline{x}_{A|B}$ is multiplied by $\frac{|B|}{|A|}$ in (9) but this does not imply a better or a worse interval than (6), the one obtained by the second sample which is not needed here: only $\overline{x}_{A|B}$ and N_B are needed to determine the confidence interval for $O(B\,|\,A)$ (provided that we know $|A|$ and $|B|$ which is easy, as explained in the previous section).

This method will now be applied in the determination of the two-fold overlap between Zentralblatt MATH and MathSciNet.

3. Application of the one-sample method to the measurement of the two-fold overlap between Zentralblatt MATH and MathSciNet

The data were collected on October 17, 2005. First of all we found that for the period 1965-2005 the size of Zentralblatt MATH was 1.988.309 and the one of MathSciNet 1.867.216 documents. These numbers are obtained via the summation (for each database) of the number of retrieved documents, year per year in the mentioned period. In order to use the formulae of the above section we hence define

A = Zentralblatt MATH

B = MathSciNet

so that $|B| < |A|$ and we will take a random sample in B. We noticed, however, that A is slower in indexing the actual publication as the following table shows. Sometimes the back-log can be nearly 2 years.

Table 1. Number of documents in A and B in the years 2005, 2004 and 2003

	A	B
2005	24.334	35.020
2004	60.562	71.492
2003	73.215	75.505

This gives interesting information concerning the indexing speed of A and B but might cause a bias when using the years 2004 and 2005 in our sample. Indeed we are interested in the overlap of both collections, once they have been put into the database. For that reason we only sampled in the period 1967-2003 and we took randomly 10 documents every two years yielding a sample of 180 documents (as said the sample was taken in B). It was quite a task to determine, manually, whether or not these documents belong to A. Indeed, difficulties were: different indexing systems in A and B and the fact that, for languages such as Russian the bibliographic index was in English (American) in B but often in German in A (certainly for the earlier years). For this reason the sample size was limited to 180. As mentioned above, the sample size determination is not the main issue of this paper. It will turn out that the obtained confidence intervals are reasonably limited. The main point of the paper is that one overlap confidence interval can be determined by the other one as we will show below.

After taking the necessary steps to overcome these problems we arrived at the following confidence interval for $O(A|B)$, i.e. the overlap of Zentralblatt MATH with respect to MathSciNet: $N_B = 180$ and we found 152 documents (out of the 180) in Zentralblatt MATH, hence $\bar{x}_{A|B} = 0.844$, yielding the 95% confidence interval

$$O(A|B)\text{Î } [0.791, 0.897] \qquad (10)$$

hence 0.844 ± 0.053 (95% sure).

Since $|A| = 1.988.309$ and $|B| = 1.867.216$, as indicated above, we hence find

$$\frac{|B|}{|A|} = 0.939 .$$

Without performing a new sample in A we can immediately conclude, by (9), using the above sample in B that, 95% sure (by (10))

$$O(B|A)\text{Î } [0.791x0.939, 0.897x0.939] = [0.742, 0.842] \qquad (11)$$

i.e. 0.792 ± 0.050.

Both intervals can be considered satisfactory based on the small sample of size $N_B = 180$, hence solving the estimation of $O(A|B)$ and $O(B|A)$ by one sample (in B).

These overlap results are elements in the decision whether or not subscriptions to both collections are continued. Of course, not only the overlap results are playing a role in this decision: also the above mentioned "speed" of indexation in A and B plays a role. For other aspects of resource selection, see [12].

4. Conclusions and remarks

In this paper we showed that one sample is sufficient to estimate both relative overlaps $O(A|B)$ (overlap of a set A with respect to B) and $O(B|A)$ (overlap of a set B with respect to A). We advise to take the sample in the smallest set (e.g. B) and determine the fraction $\bar{x}_{A|B}$ of documents in this sample that also belongs to A. Based on this fraction the usual confidence intervals for $O(A|B)$, based on the Central Limit Theorem for averages (fractions are averages), can be calculated. If $[a,b]$ is such a confidence interval for $O(A|B)$ we showed that

$$\left[a\frac{|B|}{|A|}, b\frac{|B|}{|A|}\right] \tag{12}$$

is a corresponding confidence interval (with the same probability) for $O(B|A)$. Hence no sample in A is necessary to estimate $O(B|A)$.

We then apply this theory to measure the two-fold overlap between Zentralblatt MATH and MathSciNet. We show, with one sample in MathSciNet, that the overlap of Zentralblatt MATH with respect to MathSciNet is 0.844 ± 0.053 (95% sure) and that the overlap of MathSciNet with respect to Zentralblatt MATH is 0.792 ± 0.050 (95% sure).

If we have two sets A and B as above, it is in general not possible to sample in $A \cup B$. This is so because, usually, one cannot merge two libraries or databases (e.g. the merging of MathSciNet and Zentralblatt MATH is virtually impossible). If it is possible to sample in $A \cup B$ then we can determine the number of documents in this sample that belong to $A \cap B$. Hence we can determine the fraction $\bar{x}_J$ which is the estimated value for

$$J = \frac{|A \cap B|}{|A \cup B|} \tag{13}$$

being the Jaccard index (see e.g. [13], [4], [14], [5], [15], [16] and [17]) (in [8] J is also called "total overlap"). Since $\bar{x}_J$ is a fraction we can again determine a confidence interval for J, e.g. for 95% certainty:

$$J \in \left[\bar{x}_J - 1.96\sqrt{\frac{\bar{x}_J(1-\bar{x}_J)}{N-1}},\ \bar{x}_J + 1.96\sqrt{\frac{\bar{x}_J(1-\bar{x}_J)}{N-1}} \right] \qquad (14)$$

where N is the sample size of the sample in $A \cup B$.

Note finally, that, if $A = \text{ret}$ and $B = \text{rel}$ (the set of retrieved documents, respectively the set of relevant documents in an IR process) $O(A|B)$ is nothing else than recall and $O(B|A)$ is the well-known measure precision (see [15], [17]). In fact, all the other measures (like fallout, miss – see [18]) can be described via the general relative overlap $O(A|B)$ (or $O(B|A)$) by using ret, rel, and their complements for the sets A and B.

5. References

[1] W.Y. Arms (1973). Duplication in union catalogues. *Journal of Documentation* 29(4), 373-379.

[2] M.K. Buckland, A. Hindle and P.M. Walker (1975). Methodological problems in assessing the overlap between bibliographical files and library holdings. *Information Processing and Management* 11, 89-105.

[3] L. Egghe (2006). Properties of the n-overlap vector and n-overlap similarity theory. *Journal of the American Society for Information Science and Technology* (to appear).

[4] L. Egghe and R. Rousseau (1990). *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science.* Elsevier, Amsterdam, the Netherlands.

[5] D.A. Grossman and O. Frieder (1998). *Information Retrieval Algorithms and Heuristics*. Kluwer Academic Publishers, Boston, MA, USA.

[6] L. Egghe and R. Rousseau (2001). *Elementary Statistics for effective Library and Information Service Management*. Aslib imi, London, UK.

[7] M. Gluck (1990). A review of journal coverage overlap with an extension to the definition of overlap. *Journal of the American Society for Information Science* 41(1), 43-60.

[8] W.W. Hood and C.S. Wilson (2003). Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology* 54(12), 1091-1103.

[9] A. Giral and A.G. Taylor (1993). Indexing overlap and consistency between the Avery Index to Architectural Periodicals and the Architectural Periodicals Index. *Library Resources & Technical Services* 37(1), 19-44.

[10] N. Pravdic and V. Oluić-Vuković (1987). Application of overlapping technique in selection of scientific journals for a particular discipline – methodological approach. *Information Processing & Management* 23(1), 25-32.

[11] P. Vinkler (1997). Direct and indirect literature overlap measures for information pools of research teams. *Journal of Information Science* 23(6), 433-443.

[12] S. Wu and F. Crestani (2002). Multi-Objective Resource Selection in Distributed Information Retrieval. In: *Proceedings of IPMU 02, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Annecy, France,* July 2002, 1171-1178.

[13] B.R. Boyce, C.T. Meadow and D.H. Kraft (1995). *Measurement in Information Science.* Academic Press, New York;

[14] L. Egghe and R. Rousseau (2006). Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. *Information Processing and Management* 42, 106-120, 2006.

[15] G. Salton and M.J. Mc Gill (1987). *Introduction to modern Information Retrieval.* Mc Graw-Hill, New York, NY, USA.

[16] J. Tague-Sutcliffe (1995). *Measuring Information. An Information Services Perspective.* Academic Press, New York, NY, USA.

[17] C.J. Van Rijsbergen (1979). *Information Retrieval* (2nd ed.). Butterworths, London, UK.

[18] L. Egghe (2004). A universal method of information retrieval evaluation: the "missing" link M and the universal IR surface. *Information Processing and Management* 40(1), 21-30.