Thoughts about the Stacey & Stacey measures for the logical quality of web searches

Ronald Rousseau
KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, 8400 Oostende, Belgium, & Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium & Antwerp University, IBW, Universiteitsplein 1, 2610 Wilrijk, Belgium
Email: ronald.rousseau@khbo.be

Abstract

The Stacey & Stacey formula for the quality of AND web queries is analyzed and explained, leading to a proposal for a simplified alternative. Examples clarify the procedure.

Introduction

In their book "*Effective information retrieval from the Internet* "Alison and Adrian Stacey (2004) propose a procedure and a measure for determining if search terms work together in narrowing down a search. The described procedure makes use of queries written in conjunctive form, this is: written as a sequence of ANDs. Each of the search terms is then called a conjunct. Hence, we will study queries of the form

$$Q_1 \text{ AND } Q_2 \text{ AND } Q_3 \text{ AND } \ldots \text{ AND } Q_n \qquad (1)$$

The natural number n is finite, and usually small. A conjunct may be of any query type: a disjunction (ORs), a negation (NOTs), a field search (e.g. site:uk, or inurl:Microsoft), or any other type of query possible in a search engine (Hock, 2004). Recall that the conjunctive form is the default form in many search engines, including Google.

Each conjunct in a query returns a set of hits. The set returned by a query Q will be denoted as S(Q). Clearly, the set returned by Q_1 AND Q_2, is the intersection of the sets returned by Q_1 and Q_2 separately:

$$S\left(Q_1 \ AND \ Q_2\right) = S(Q_1) \cap S(Q_2) \qquad (2)$$

Beginning with Q_1 and the corresponding set $S(Q_1)$ one may say that $S(Q_2)$, the set returned by query Q_2 , cuts through $S(Q_1)$, leading to $S(Q_1) \cap S(Q_2)$. The

point is that if $S(Q_1) \cap S(Q_2)$ is only slightly smaller than $S(Q_1)$, the two search terms do not work together efficiently in order to narrow down the search. An efficient web search is defined here as one where all the intersections in expression (1) considerably reduce the preceding result. The smaller the new intersection (in relative terms) the more efficient we say that the search is. Or, expressed differently, the better the terms collaborate or 'join forces' in reaching the goal, which is a relatively small set of highly relevant sites, the more efficient the search is. Note though that there is one important caveat. If $S(Q_1) \cap S(Q_2) = \emptyset$ then the search has not reached its goal (is not effective anymore). In this case search terms Q_1 and Q_2 do not collaborate but exclude one another. Hence we try to reduce $S(Q_1) \cap S(Q_2)$ to a small set, without 'overdoing' it.

Purpose

Stacey & Stacey propose a measure for the efficiency of search terms in a search on the Web. For a query consisting of two terms they propose the TTQQM: Two-Term Quality of Query Measure. The calculation of the TTQQM will be explained further, but basically it is just an overlap measure. For queries consisting of more terms they propose a generalization of TTQQM, using an averaging procedure. In this article we will explain these procedures, propose an alternative, and perform an experiment.

We would like to point out that we perform an evaluation of searches from a logical perspective. The quality of the obtained information is not evaluated, nor is the efficiency of a search engine's technology.

The Stacey & Stacey TTQQM measure

Consider a search consisting of two search terms: Q_1 and Q_2. The Stacey & Stacey two-term quality of query measure (in short: TTQQM) is calculated as follows:

$$TTQQM(Q_1, Q_2) = 1 - 2\ \frac{\#S(Q_1 \bigcap Q_2)}{\min\left(\#S(Q_1), \#S(Q_2)\right)}$$

Here #S refers to the number of elements in the set S. In practice Stacey & Stacey propose the following procedure:

- Search for Q_1 by itself and for Q_2 by itself. From now on the one with the larger number of hits is called Q_2, the other one Q_1.
- Now search for Q_1 combined with itself. The number of hits is denoted as s.

- Now search for Q_1 combined with Q_2, i.e. Q_1 AND Q_2 and, separately, Q_2 combined with Q_1, i.e. Q_2 AND Q_1. These results should be the same but in reality this is often not the case. The average number of hits of these two searches is called c.
- TTQQM = $1 - \frac{2c}{s}$.

The idea behind this practical procedure is that one starts from the most focussed query, this is: the one returning the smaller number of hits. Then one combines this query with another query in order to obtain an even smaller number of hits. The worst possible way to do this is using the same query again, as then there is no reduction at all. In practice, see examples, Q AND Q does not return the same number of hits as Q (although it must theoretically be the same) so Stacey & Stacey recommend using the results of the query Q AND Q. Similarly, the queries Q_1 AND Q_2 and Q_2 AND Q_1 should yield the same results, but again, in practice they often differ somewhat. For this reason they recommend using the average of the two.

Comments on the procedure for determining Stacey & Stacey's TTQQM.

A. TTQQM always yields a value between -1 and +1. Indeed, the number c is at least 0 (no overlap) and at most s (the set retrieved by Q_1 is a subset of the set retrieved by Q_2). So clearly $-1 \leq TTQQM \leq 1$.

B. When half of the items retrieved by Q_1 AND Q_2 belongs to the set retrieved by Q_1 (actually Q_1 AND Q_1) then TTQQM = 0. Recall also that in practice TTQQM must be strictly smaller than one, otherwise no items are found.

C. According to Stacey & Stacey, a TTQQM in excess of 0.2 is considered to be indicative that the searcher chose the two search terms in such a way that they worked well in conjunction with each other. Note that this is just a rule of thumb.

D. TTQQM is nothing but a renormalization of the basic overlap measure $O_1(Q_1,Q_2) = \frac{\#S(Q_1 \cap Q_2)}{\min\left(\#S(Q_1),\#S(Q_2)\right)}$ (Salton & McGill, 1983, p.203; Egghe & Michel, 2002). Indeed, when O_1 is 1 then TTQQM is -1, when O_1 is 0.5 then TTQQM is 0, and when O_1 is 0 then TTQQM is 1.

E. Egghe and Michel (2002) study properties of similarity measures and find that $O_2(Q_1,Q_2) = \frac{\#S(Q_1 \cap Q_2)}{\max\left(\#S(Q_1),\#S(Q_2)\right)}$ has better properties than O_1. In their terminology O_2 is a strong similarity measure, while O_1 is only a weak similarity measure. Yet, because we have a different purpose, namely the measurement of the focus of a search, we follow Stacey & Stacey and prefer O_1 for our study.

An example (search performed on January 31, 2006, in Google)

We would like to find the site of the Science and Technology Indicators Conference in Leuven (2006). This conference has been organized before by CWTS in Leiden. For this reason we used the search terms Leuven and CWTS. Searching for *CWTS* yielded 350,000 hits which turned out to be equal to the number of hits obtained by *CWTS AND CWTS*. The number of hits for *Leuven* was 19,200,000. *CWTS AND Leuven* resulted in 236 hits, while *Leuven AND CWTS* gave 425 hits. This is quite a difference!

Anyway these two search terms work well together (information about the Science & Technology Indicators Conference came at ranks 4 and 3, respectively). Their TTQQM is 0.9981 (using an average value for the intersection).

The Stacey & Stacey generic quality of query measure

For queries consisting of more than two search term Stacey & Stacey (2004) propose a somewhat different procedure, which will be explained by considering the case of four terms.

Let us consider a query consisting of the four search terms: Q_1 AND Q_2 AND Q_3 AND Q_4. Which is the one considered to be added the last? It is assumed that the last one is the term which has the worst effect on the conjunction of the previous three. Hence, one considers all groups of three terms (for a general n-term query, this is all possible groups of n-1 terms). In general there are $\binom{n}{n-1}$ = n possible cases. For n = 4 there are 4 cases, namely (Q_1 AND Q_2 AND Q_3), (Q_1 AND Q_2 AND Q_4), (Q_1 AND Q_3 AND Q_4), and finally (Q_2 AND Q_3 AND Q_4). The missing term of the group with the smallest number of hits is the one considered to be added last. Let us assume that it is Q_4. This procedure is repeated for the three remaining query terms, and finally also for the two remaining ones. Assume that the order in which terms are added is given as Q_1, Q_2, Q_3, Q_4.

If c denotes the number of hits for Q_1 AND Q_2 AND Q_3 AND Q_4 and s the number of hits for Q_1 AND Q_2 AND Q_3 then the term quality of the last term, denoted as SS_4, is just the complement of the overlap measure O_1 : $SS_4 = 1-\frac{c}{s}$. Note that Stacey & Stacey actually define c as the maximum number of hits obtained from the three queries (Q_1 AND Q_2 AND Q_3 AND Q_3), (Q_1 AND Q_2 AND Q_2 AND Q_3) and (Q_1 AND Q_1 AND Q_2 AND Q_3). We would assume that if these are so different that it really matters then the whole procedure of determining a quality measure would be rather futile. Hence we will not do this.

This procedure is repeated for the smaller query Q_1 AND Q_2 AND Q_3 leading to SS_3 and finally for Q_1 AND Q_2 leading to SS_2. In general, for an n-term query one obtains n-1 SS-measures. Note that all SS-measures take values between 0 and 1 and are, at this moment, not normalized to yield values between -1 and +1 as is TTQQM.

Next, Stacey & Stacey propose a rather ad hoc normalization. They state that a threshold value is necessary for discriminating good from bad queries, and propose 0.6 as such a value. Then they consider the parabola through the points (0,-1), (0.6, 0) and (1, 1). In this way one obtains the equation $0.833\ x^2 + 1.167\ x - 1$. Substituting an SS-measure for x yields the standardized term quality.

Finally the quality of the search is obtained by taking the (arithmetic) average of the standardized term quality measures.

We present an example in order to illustrate this procedure. Suppose we want to find information about vaccines against bird flu in China or from a Chinese perspective.

Let Q_1 be the query site:cn, Q_2 the query (vaccin OR vaccine) and Q_3 the query "bird flu" in Google (search performed on February 4, 2006). The order of these queries was determined from

site:cn AND (vaccin OR vaccine): 108,000 hits
site:cn AND "bird flu": 159,000 hits
(vaccin OR vaccine) AND "bird flu" 1,800,000 hits

From these three results we conclude that the search term "bird flu" must play the role of Q_3.

Next we perform the queries:
site:cn : 61,900,000 hits
(vaccin OR vaccine) : 38,600,000 hits

From these results we derive that (vaccin OR vaccine) is Q_1 and site:cn plays the role of Q_2.

The complete query: site:cn AND (vaccin OR vaccine) AND "bird flu" yields 21,400 hits, hence $SS_3 = 1 - \frac{21{,}400}{108{,}000} = 0.802$. As the search (vaccin OR vaccine) AND (vaccin OR vaccine) yields 38,500,000 hits, $SS_2 = 1 - \frac{108{,}400}{38{,}500{,}000} = 0.997$.

The normalized forms of these SS-measures are obtained from the equations:

$$0.833\ (0.802)^2 + 1.167(0.802) - 1 = 0.472 \text{ and}$$

$$0.833\ (0.997)^2 + 1.167(0.997) - 1 = 0.992$$

The final quality of this search is the average of these two normalized measures: (0.472+0.992)/2 = 0.732. This is a very good result, although we still have 21,400 hits! This is no surprise as the query itself was rather vague.

An alternative, slightly simplified approach

Instead of the elaborate procedure and the ad hoc threshold proposed by Stacey and Stacey we would like to propose the following alternative.

If a query consists of the terms Q_1, Q_2, Q_3 AND Q_4 we suggest searching for each of these query terms separately and ranking them according to the number of hits they receive, starting from the one with the least number of hits. Assume that Q_1 is the query with the least number of hits, followed by Q_2, Q_3 and finally Q_4, retrieving the largest set.

Calculate now

$$T_1 = 1 - \frac{\#S(Q_1\ AND\ Q_2)}{\#S(Q_1)},$$

$$T_2 = 1 - \frac{\#S(Q_1\ AND\ Q_2\ AND\ Q_3)}{\#S(Q_1\ AND\ Q_2)},$$

$$T_3 = 1 - \frac{\#S(Q_1\ AND\ Q_2\ AND\ Q_3\ AND\ Q_4)}{\#S(Q_1\ AND\ Q_2\ AND\ Q_3)}$$

In general, for an AND-query consisting of k conjuncts k-1 T-terms are calculated, according to the formula

$$T_j = 1 - \frac{\#S\left(Q_1\ AND \cdots AND\ Q_{j+1}\right)}{\#S\left(Q_1\ AND \cdots AND\ Q_j\right)},\quad j = 1, \ldots, k-1$$

where queries are ranked according to the number of hits each retrieves separately. Ranking occurs from smallest to largest. This fixed ranking is also meant to take care of the problem that sometimes Q_1 AND Q_2 and Q_2 AND Q_1 do not retrieve the same number of hits. Note that the proposed order in which queries are considered is just a heuristic device. There is no claim that this order is optimal in some sense.

Finally a simple arithmetic or geometric average yields the new QuaLity measures for AND-queries, denoted as QLA_a in the case an arithmetic average is taken and QLA_g when a geometric average is used:

$$QLA_a = \frac{1}{k}\sum_{j=1}^{k} T_j \qquad\qquad QLA_g = \sqrt[k]{\prod_{j=1}^{k} T_j}$$

For a 4-term query the number of searches to be done in the S&S-procedure is 4 + 3 + 2 (and one more if the Q_1 AND Q_1 query is added). In general this is n (n-1)/2 queries for an n-term query. Then 2 times 3 +1 (in general: 2 times (n-1) +1) calculations must be performed.

In the new procedure one starts with n queries, and then exactly one query for each level (number of query terms). This yields n + (n-1) = 2n -1 searches to be performed. Then n calculations must be performed (no normalization is necessary). For n small this does make a serious difference, but for e.g. an elaborate 8-term query the first procedure needs 28 searches while the second one only needs 15 searches (and fewer calculations).

When the geometric average is used a zero value indicates that one of the T-values is zero, which happens when for some j

$$\#S\left(Q_1\ AND \cdots AND\, Q_j\ AND\ Q_{j+1}\right) = \#S\left(Q_1\ AND \cdots AND\ Q_j\right).$$

This equality means that adding term Q_{j+1} did not reduce the retrieved set. Thus Q_{j+1} is a completely superfluous term.

An example

We searched for the ten articles published in the journal *Research Evaluation* vol.14 (2), August 2005 (see Table of Contents in the Appendix).

We used the surname of the first author and three expressions taken from the title. Searches were performed in Google and a comparison was taken between the results obtained by the Stacey & Stacey algorithm and the simplified one. The exact queries and the resulting number of hits are shown in Table 1. When Google showed a result as, e.g. *1-4 of about 6* this was counted as 6. Otherwise this would have resulted in serious irregularities.

Table 1. Queries and number of hits (search performed in February 2006)

Query: conjuncts are shown	Number of hits
Granadino MCYT "Acciones Integradas" "Spanish scientific output"	4
Jin "Key labs" "Open labs" Chinese	9
Costas "bibliometric indicators" "natural resources" CSIC	5
Newman "28 nations" decade "competitive performance"	6
Esterle France "public research organisations" unique	22
Gomez regionalisation "science and technology data" Spain	1
Antonangeli "social accountability" Elettra project	7
Grohmann German 1990s "on-line bibliometric analysis"	4
Sigogneau "cross-disciplinary research" CNRS practices	8
"Modrego-Rico" indicators measure performance	17

The searches we performed are very specific searches and we expected them to end up with the exact results, which they did. This was no surprise as the IngentaConnect website contains the table of contents of *Research Evaluation*. Note also that we did not try to find the best possible or the most efficient search formulation, but just wanted to illustrate the way in which the two types of measures work in actual web searches.

Table 2 shows that there is no essential difference between our simple procedure and the more elaborate and time-consuming procedure proposed by Stacey and Stacey. Note that the Stacey and Stacey procedure yields values between -1 and +1, while ours gives values between 0 and 1. In both cases the higher value corresponds with the better (= more efficient) search.

Table 2. Comparison between the Q&Q-measure and QLA_a

Name of search: first author	Value of Q&Q measure	QLA_a
Costas	0.344	0.809
Gomez	0.266	0.643
Modrego-Rico	-0.034	0.559
Jin	-0.011	0.528
Esterle	-0.042	0.514
Sigogneau	-0.173	0.452
Antonangeli	-0.306	0.356
Granadino	-0.386	0.314
Newman	-0.370	0.300
Grohmann	-0.667	0.200

Differences between the two rankings are clearly small (the Spearman rank correlation coefficient is equal to 0.98) and due to two factors: the proposed simplification (the position of the article by Granadino et al.) and the fact that the average of a quadratic normalization is not equal to the quadratic normalization of an average (the position of the Jin et al. article). Note also that because these searches are very focused (the title of one particular scientific article, where the first author is one of the search terms), they are generally rather poor, in the sense that very similar search results can be obtained with three or even two search terms. As we always used four search terms, as an illustration of the procedure and its meaning, this resulted in rather inefficient searches. Recall that the term 'inefficient' does not refer to the final result, but to the way adding new search terms reduces the number of results of the preceding result.
The inefficiency of the used procedure becomes very clear when considering the QLA_g. Most of these values are equal to zero, indicating that at some step in the procedure no progress had been made (see Table 3). Table 3 also shows that the first step, i.e. going from one search term to two usually leads to an enormous reduction in the size of the retrieved data set. T_1- data are rounded to three decimals, except when the number is larger than 0.999; then they are rounded to four decimals.

Table 3. QLA_g and T_1-values

Name of search: first author	QLA_g	T_1
Costas	0.757	0.986
Gomez	0	0.999
Modrego-Rico	0.494	0.876
Jin	0	0.998
Esterle	0	0.9992
Sigogneau	0.289	0.999
Antonangeli	0	0.9997
Granadino	0	0.943
Newman	0	0.9998
Grohmann	0	0.600

Further theoretical considerations

It seems that in practice T_1 is larger than T_2 which is larger than T_3 (if there are four search terms). Yet, there is no mathematical reason why this should be the case. We provide an example where the opposite inequalities $T_1 < T_2 < T_3$ hold. Consider the query results shown in Fig.1. We note that it is not entirely trivial to find a configuration showing all possible intersections of four sets (Rousseau, 1998).

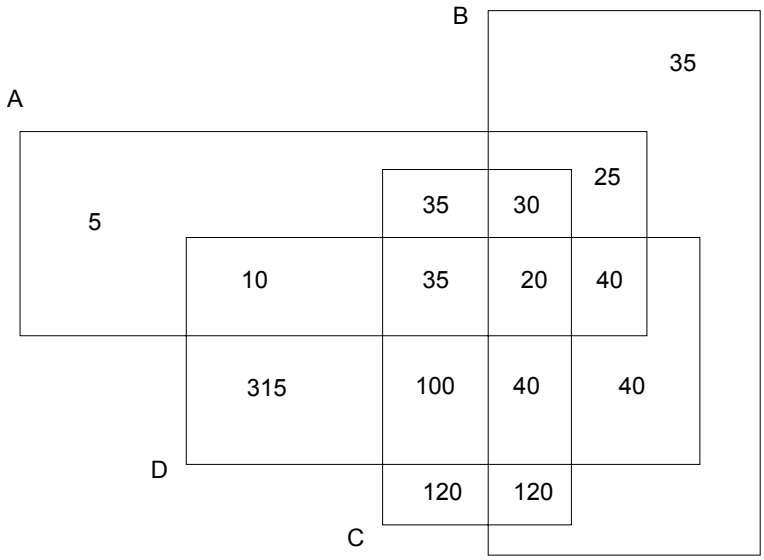

Fig.1 Fictitious example of retrieved sets of four queries and numbers of retrieved items in each subset

Let A = S(Q_1), B = S(Q_2), C = S(Q_3) and D = S(Q_4). Then #A = 200, #B = 350, #C = 500 and #D = 600. In this way sets are already ranked correctly according to the simplified version. Then

$$T_1 = 1 - \frac{115}{200} = 0.425$$

$$T_2 = 1 - \frac{50}{115} = 0.565$$

$$T_3 = 1 - \frac{20}{50} = 0.600$$

The corresponding QLA_a and QLA_g-values are 0.53 and 0.525.

In order to calculate the SS-measures we have to determine: $\#(A \cap B \cap C) = 50$, $\#(A \cap B \cap D) = 60$, $\#(A \cap C \cap D) = 55$ and $\#(B \cap C \cap D) = 60$. Hence the last set in the SS-procedure is D. Hence, C is the third set. Finally, #A = 200 and #B = 350 leading the same order as for the simplified procedure. Consequently the SS-measures are the same as the T-measures. This shows that also for the Stacey & Stacey approach the values may occur in decreasing order. We consider this a good property. Indeed, if

consecutive T- or SS-values always decreased, adding new search terms would always decrease the global efficiency of a query as measured by the Stacey & Stacey approach. If that had been the case then the Stacey & Stacey approach would not really be interesting.

Conclusion and ideas for further research

We explained and illustrated the use of the Stacey & Stacey measures for the efficiency of web searches. Moreover, we proposed a simplified procedure. Indeed, we do not think that a general approach (a mathematical formula) should take search engines' idiosyncrasies into account and make provisions for differences between the number of hits retrieved by Q_1 AND Q_2 and those retrieved by Q_2 AND Q_1. Of course, in practice this might be a sensible thing to do if there is indeed a large difference between the two.

The Stacey & Stacey measures, in original or in simplified form, lead, in our opinion, to a promising approach for comparing the logical efficiency of different search engines. These measures may also be used to compare searches on different topics, where these searches may be performed with the same search engine or using different ones. Furthermore, different types of search terms, e.g. phrases vs. single words, may be compared using the S&S measures. We expect to accomplish at least part of this research program in the near future. Finally, we hope that our article may lead to further research in comparative evaluation of search engines.

Acknowledgements. The author thanks Gwendolyn Rogge (KHBO) for editorial help. Research for this article began when the author visited Henan Normal University and the National Library of Sciences CAS (Beijing). He thanks Prof. Liang Liming, Prof. Jin Bihui, their colleagues and students for their hospitality. He also acknowledges support from the NSFC Grant Nr. 70373055.

References

Egghe, L. and Michel, C. (2002). Strong similarity measures for ordered sets of documents in information retrieval. *Information Processing and Management*, 38, 823-848.

Hock, R. (2004). The latest field trip: an update on field searching in Web search engines. *Online*, 28(5), 15-21.

Rousseau, R. (1998). Venn, Caroll, Karnaugh and Edwards. *Wiskunde & Onderwijs*, 24, 233-244.

Salton G. and McGill, Michael J. (1983). *Introduction to modern information retrieval*. Singapore: McGraw-Hill.

Stacey, A. and Stacey, A. (2004). *Effective information retrieval from the Internet*. Oxford: Chandos Publishing.

Appendix

Table 1: Research Evaluation, volume 14(2), August 2005