Stopping Criteria for Technology Assisted Reviews based on Counting Processes

Alison Sneyd and Mark Stevenson
University of Sheffield
mark.stevenson@sheffield.ac.uk

ABSTRACT

Technology Assisted Review (TAR) aims to minimise the manual judgements required to identify relevant documents. Reductions in workload are dependent on a reviewer being able to make an informed decision about when to stop examining documents. Counting processes offer a theoretically sound approach to creating stopping criteria for TAR approaches that are based on analysis of the rate at which relevant documents are observed. This paper introduces two modifications to existing approaches: application of a Cox Process (a counting process which has not previously been used for this problem) and use of a rate function based on a power law. Experiments on the CLEF 2017 e-Health TAR collection demonstrates that these approaches produces results that are superior to those reported previously.

CCS CONCEPTS

• **Information systems** → **Retrieval effectiveness**; **Retrieval efficiency**.

KEYWORDS

Technology assisted review; TAR; total recall; stopping criteria; counting processes; Cox Process; Poisson Process; CLEF eHealth

ACM Reference Format:
Alison Sneyd and Mark Stevenson. 2021. Stopping Criteria for Technology Assisted Reviews based on Counting Processes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463013

1 INTRODUCTION

Technology Assisted Review (TAR) aims to minimise the manual effort required to assess collections of documents for relevance. Strategies such as active learning have been shown to be highly successful at ensuring that those of interest appear early in the ranking, e.g. [1, 2, 6]. However, a highly successful ranking strategy is not itself sufficient to cut down the effort required to review a collection since a reviewer needs to know when it is safe to stop examining documents. Stopping strategies, e.g. [1, 8, 13], offer a solution by predicting an appropriate point at which the process of examining documents can cease. They offer the potential to reduce the effort required for TAR by removing the requirement to examine the entire document collection, particularly those that are unlikely to be of interest (such as those that are low ranked).

Existing stopping criteria are generally based on either ad hoc approaches or methods that attempt to estimate the total number of relevant documents in the collection (see Section 2). Sneyd and Stevenson [13] introduced an alternative approach based on counting processes, stochastic processes that model the number of event occurrences within a time interval. This approach was demonstrated to be robust by evaluating it using rankings of varying effectiveness produced by participants from the CLEF eHealth task "Technology Assisted Reviews in Empirical Medicine" [6]. However, their work only explored a single counting process (Poisson Process) and assumed that the rate at which relevant documents are observed can be modelled as an exponential function. In addition, they only reported results for a limited range of parameter settings (recall of 70% with 95% confidence).

This paper extends previous work on the use of counting processes for stopping criteria in multiple ways. It applies a counting process that has not previously been applied to this problem (the Cox Process). It explores the use of alternative rate functions, specifically power laws [16], which are more commonly applied in IR research than the exponential function used previously. It also reports results for a range of recall and confidence levels.

The paper's main contributions are: to introduce the Cox Process for this problem; to demonstrate that using power laws to model the rate at which relevant documents are observed is more effective than the exponential function; to provide information about how these approaches behave under a range of recall and confidence levels and compare them against existing baselines.[1]

2 BACKGROUND

The simplest approaches to developing search stopping criteria are based on heuristics, such as stopping after a set percentage of the ranked documents have been examined or stopping after a set number of consecutive non-relevant documents have been observed, e.g. [9]. However, these approaches rely heavily on parameters being set to appropriate values which is difficult to achieve when the approach is applied to a new collection or ranking algorithm.

A range of methods for identifying search stopping criteria have been developed based on the approach of first estimating the total number of relevant documents in the collection. It is then straightforward to create a stopping criteria based on a pre-defined level of recall by simply stopping when the appropriate portion of the relevant documents have been observed. Such approaches can also be used to provide confidence estimates based on the confidence in the estimation of the total number of relevant documents. This

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

[1] Code implementing the experiments described in this paper can be found at https://github.com/alisonsneyd/stopping_criteria_counting_processes

approach is very closely related to work on volume estimation e.g. [10, 15], where the goal is to estimate the total number of relevant documents in a (possibly unranked) collection.

One stopping method based on this approach, the *target method* [1], was motivated by the problem of electronic discovery in the legal domain. The target method involves examining randomly chosen documents until a set number of relevant ones have been found; these are referred to as the target set. The method then proceeds through the ranking, making judgements until every document in the target set has been encountered. The size of the target set depends on the level of recall and confidence in this level that is desired. A target set of 10 is sufficient to guarantee a minimum recall score of 0.7 with probability 0.95, see [1]. The target method does not make any assumptions about the distribution of relevant documents in a ranking. For a reasonable retrieval system relevant documents are more likely to occur at the start of the ranking than at the end, but the target method does not exploit this information.

The knee method [1] aimed to overcome this limitation by assuming that the gain curve of rank versus recall is convex in shape. It employs geometric analysis [12] to estimate when the gain curve has exceeded a certain tipping point, called the slope ratio. This method was better able to predict an efficient stopping point than the target method, but did not provide any probabilistic guarantees about the level of recall that would be achieved.

Other approaches include Wallace et al. [14] who proposed a number of measures to estimate the total number of relevant documents in a collection and provided this information to the user to allow them to make the decision about when to stop. Recently Li and Kanoulas [8] also developed estimates, which they proved to be statistically unbiased, and integrated them into an active learning approach to document review.

Sneyd and Stevenson [13] proposed an alternative approach based on analysis of the rate at which relevant documents are identified in a ranking. Their approach examines the documents highest in the ranking (i.e those most likely to be relevant) and models the rate at which relevant documents are likely to be found in the examined portion of the ranking using a statistical model. They chose to model the rate at which relevant documents are found in the ranking used an exponential function. The approach was demonstrated to be robust on a set of rankings of varying effectiveness produced by participants in a shared task [6]. Since it is based on an established statistical model, this approach has the advantage of providing estimates of the total number of relevant documents in a ranking together with confidence estimates. However, Sneyd and Stevenson do not justify the use of an exponential function to model the rate at which relevant documents occur and also point out that it might not be suitable in all circumstances.

3 APPROACH

This section describes how counting processes can be used to produce stopping rules, building on previous work [13]. Counting processes are stochastic models of the number of occurrences of an event over time (see e.g. [5]). They can naturally be applied to the problem of estimating the number of relevant documents in search results by treating positions in a search ranking as “time” and occurrences of relevant documents within this ranking as “events".

Poisson processes [7] assume that the events occur independently and that the number of occurrences in given interval follows a Poisson distribution. If the rate at which events occur is constant, the Poisson process is called homogeneous. If the rate varies then the Poisson process is inhomogeneous (or non-homogeneous). These Poisson processes have a rate, λ, which is a function representing the frequency with which events occur over the space over which the process is defined (in our case this space is a ranked list of documents). Let:

$$\Lambda(a,b) = \int_a^b \lambda(x)dx, \tag{1}$$

and let $N(a,b)$ be a random variable denoting the number of events an occurring in the interval $(a,b]$. Then:

$$P\left(N(a,b) = r\right) = \frac{[\Lambda(a,b)]^r}{r!} e^{-\Lambda(a,b)}. \tag{2}$$

N has a Poisson distribution with expected value $\Lambda(a,b)$. This provides a mechanism to model the number of relevant documents found between indexes a and b of a ranking and estimate the probability of this being a particular value (r). Restricting attention to the interval $N(0,n)$, where n is the total number of documents in the collection, provides a distribution that can be used to estimate the number of relevant documents in the entire collection.

Selecting an appropriate rate function is important to ensure the Poisson process represents an accurate model, but in most cases the rate function is not known and cannot be directly observed. Cox processes [4], also known as doubly stochastic Poisson processes, are an extension of Poisson Processes that account for this uncertainly by also modelling the rate function as a stochastic process. In a Cox process, the probability that a defined number of events, r, occurs within the interval $(a,b]$ is computed by integrating over all of the possible values of the rate function λ, i.e.

$$P\left(N(a,b) = r\right) = \int_0^\infty \frac{[\Lambda(a,b)]^r}{r!} e^{-\Lambda(a,b)} P(\Lambda) d(\lambda). \tag{3}$$

3.1 Estimating the Rate Function

In this application the rate function, λ, represents the probability of a document being relevant at a particular point in the rankings and choosing a suitable one is a key decision in the application of counting processes. An appropriate function should assume that a suitable ranking has, in accordance with the probability ranking principle [11], succeeded in placing documents that are more likely to be relevant higher in the ranking than those less likely to be and, consequently, the rate at which relevant documents occur decreases in direct proportion to the document's position in the ranking.

Previous work on the application of counting processes to stopping [13], chose to model λ using an exponential function, i.e. $\lambda(x) = de^{kx}$ where x is an index in a ranking (i.e. $x \in \{1, 2 \ldots n\}$ for a collection of n documents) and $d, k \in \mathbb{R}$ are parameters controlling the function's shape. This choice of rate function has mathematical properties that make it a convenient choice of rate function for a counting processes (see [13]), but it is unclear whether it is a suitable model for representing the occurrence of relevant documents.

Power laws have been proposed as a suitable model of the rate at which relevant documents are found in a ranking [16] and have been shown to be useful for estimating the number of relevant documents remaining for test collection development, e.g. [9]. Fortunately the mathematical properties of these functions also make them a convenient choice of rate function for counting processes. Power laws have the form $\lambda(x) = cx^k$ where x is an index and the parameters $c, k \in \mathbb{R}$ determine the function's shape. Substituting this into equation 1 produces:

$$\Lambda(a, b) = \int_a^b cx^k dx = \left[\frac{cx^{k+1}}{k+1} \right]_a^b = \frac{c}{k+1} \left(b^{k+1} - a^{k+1} \right). \quad (4)$$

Since we are interested in estimating the total number of relevant documents in the collection (n), attention can be restricted to the interval $(0, n]$, then

$$\Lambda(0, n) = \frac{cn^{k+1}}{k+1}. \quad (5)$$

This equation provides a convenient closed form expression that can be substituted into Equations 2 and 3 to estimate a distribution of the number of relevant documents in a collection.

3.2 Stopping Rule

Combining Equations 2 and 3 with a rate function allows the production of distributions estimating the number of relevant documents within some portion of the ranking. Examining this distribution's cumulative distribution function (CDF) provides an upper limit on this estimate, given a specified probability. For example, the maximum number of relevant documents in the collection with 95% probability. This allows us to develop an algorithm for determining a suitable stopping point for examining documents in the ranking with confidence bounds on the estimates produced. The proposed algorithm (see Algorithm 1) is provided with a desired level of recall (ℓ) and confidence level (p). The first α documents in the ranking are examined, and an upper bound with confidence p is placed on the total number of relevant documents in the entire ranking (denoted by R); this is estimated by computing the CDF of the counting process model being used (lines 3-6). If the number of relevant documents observed in the first α documents ($rel(\alpha)$) is greater than or equal to the number of relevant documents needed to achieve the desired level of recall (i.e. $rel(\alpha) \geq \lceil \ell R \rceil$) (line 7), the process stops and no more documents are examined. Otherwise, the next β documents are added to the pool of examined documents (line 10) and the process repeated. The algorithm continues until either enough relevant documents have been found to achieve the desired recall level or all documents have been examined.[2]

4 EXPERIMENTS

Experiments were carried out to evaluate the approaches based on counting processes described in Section 3 and compare them against existing methods (Section 2). Each method was tested at a variety of recall ($\ell \in \{0.7, 0.8, 0.9, 0.95\}$) and probability levels ($p \in \{0.8, 0.95\}$). For the counting process methods, the initial sample of the ranking to examine (α) was set to 30% and the percentage of new documents to examine at each iteration of the algorithm (β) set to 5%. In addition, at least 20 relevant documents had to be observed before a stopping point was proposed (c.f. previous approaches that required 150 relevant documents [1]).

Algorithm 1 Algorithm to Generate Search Stopping Criteria

```
1:  **Input:** n (= no. documents in ranking), $\ell$ (= target recall level,
    e.g. 0.7), p (= confidence level, e.g. 0.95), $\alpha$ (= initial sample size,
    e.g. 0.3), $\beta$ (= sample increment size, e.g. 0.05)
2:  **Output:** s (= stopping rank)
3:  $s \leftarrow \alpha \times n$
4:  **while** $s < n$ **do**
5:      Fit Counting Process (N(0, s)) to documents in range 1 ... s
6:      $R \leftarrow CDF\ of\ N(0, s) > p$
7:      **if** $\ell R < rel(s)$ **then**
8:          break
9:      **end if**
10:     $s \leftarrow s + \beta \times n$
11: **end while**
12: return $s$
```

The first approach, **Inhomogeneous Poisson Process with Exponential Rate (IP-E)**, uses a Poisson Process with an exponential function as the rate function. The parameters of the rate function were estimated using a non-linear least squares algorithm, the same methods as used in previous work [13]. The next approach, **Inhomogeneous Poisson Process with Exponential Rate (IP-E)**, is the same as IP-E except that the rate is determined using a power curve function rather than an exponential. The next two approaches, **Cox Process with Exponential Rate (CX-E)** and **Cox Process with Power Rate (CX-P)** replace the Poisson Process with a Cox Process with rates modelled as exponential and power law functions respectively. Parameters of the rate function as identified using non-linear least squares are normally distributed. Examining the covariance of each parameter allows the distribution of λ required for the Cox Process (see Eq. 3) to be computed.

For comparison, the **target method (TM)** [1] was implemented with the target set size adjusted to the relevant recall ℓ and probability p levels. The **knee method (KM)** [1] was also implemented but does not offer an intrinsic way to adjust its desired recall or confidence levels. Consequently the algorithm's parameters were set to their recommended values: slope ratio (ρ) to 6 and early stopping prevention parameter (β) to 150.

Finally, an **Oracle Method (OR)** was implemented to stop exactly when the desired recall level has been achieved. This approach requires complete information about the ranking and is clearly not feasible as a practical approach. However, it is included for comparison purposes since it shows the maximum number of documents that can remain unexamined while still achieving the desired recall.

4.1 Data

Following [13], we utilise the publicly available submissions to the CLEF 2017 e-Health Lab Task 2 "Technology Assisted Reviews in Empirical Medicine" [6].[3] Participants in the task were required to rank documents retrieved from a complex Boolean query for a set of 30 systematic reviews, called topics. The total number of documents across all topics is 117,562, with a per topic median of

[2] Note that this approach can be applied to a pre-ranked collection of documents or naturally integrated into an approach employing active learning, with the sample size parameters (α and β) adjusted as required.

[3] Available from https://github.com/CLEF-TAR

	Mean Recall (↑)						**Effort (↓)**							**PES (↑)**						
ℓ	TM	KM	IP-E	CX-E	IP-P	CX-P	TM	KM	IP-E	CX-E	IP-P	CX-P	OR	TM	KM	IP-E	CX-E	IP-P	CX-P	OR
p = 0.95																				
0.95	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	102,105	86,243	116,521	117,558	63,605	**63,314**	22,637	13.1	26.6	0.9	0.0	45.9	**46.1**	80.7
0.90	0.98	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	83,954	86,243	103,509	99,314	**61,361**	61,711	16,706	28.6	26.6	12.0	15.5	**47.8**	47.5	85.8
0.80	0.95	**1.00**	0.99	0.99	**1.00**	**1.00**	62,867	86,243	73,728	80,040	**56,840**	57,455	9675	46.5	26.6	37.3	31.9	**51.7**	51.1	91.8
0.70	0.92	**1.00**	0.97	0.98	0.99	0.99	55,005	86,243	54,754	**48,290**	53,585	53,263	7,419	53.2	26.6	53.4	**58.9**	54.4	54.7	93.7
p = 0.8																				
0.95	0.99	**1.00**	**1.00**	0.99	**1.00**	**1.00**	86,381	86,243	95,235	92,819	**59,703**	60,043	22,637	26.5	26.6	19.0	21.0	**49.2**	48.9	80.7
0.90	0.96	**1.00**	0.99	0.99	**1.00**	**1.00**	66,476	86,243	79,563	83,510	**58,619**	59,113	16,706	43.5	26.6	32.3	29.0	**50.1**	49.7	85.8
0.80	0.90	**1.00**	0.98	0.98	0.99	0.99	**52,417**	86,243	56,051	56,050	54,948	54,758	9675	**55.4**	26.6	52.3	52.3	53.3	53.4	91.8
0.70	0.85	**1.00**	0.97	0.97	0.99	0.99	**43,396**	86,243	46,783	47,622	51,479	51,236	7,419	**63.1**	26.6	60.2	59.5	56.2	56.4	93.7

Table 1: Results of experiments described in Section 4. ↑ and ↓ show metrics where higher and lower values are preferred (respectively). Highlighted results indicate the best non-Oracle approach.

2,070 documents and 38 relevant documents (the relevance judgements provided at abstract level are used in this paper). On average over the topics, only 1.58% of documents were relevant. The most effective set of rankings submitted to the task, i.e. the run with the highest AURC (area under recall curve) score, were used for the experiments; WaterlooB-rank-normal [3].

4.2 Evaluation Metrics

Stopping criteria aim to balance two opposing goals: retain a minimum level of recall, whilst also minimising the number of documents examined. The following metrics quantify these objectives and have been used in previous work on stopping criteria [1, 8, 13]: Given a stopping criteria S and topic (document set) T such that E_T is the set of documents in T that were examined before stopping, the **effort** of S for T is defined as $|E_T|$. Extending this definition, the effort of S for a run $\mathcal{R}$ (a set of topics ranked by the same system) is given by $\sum_{T \in \mathcal{R}} |E_T|$. We also compute the **percentage of effort saved** by S over $\mathcal{R}$ as the percentage of all documents in the run that are not examined:

$$100 \times \sum_{T \in \mathcal{R}} \frac{|T| - |E_T|}{|T|}.$$

The (macro-averaged) **mean recall** of a run over each topic is reported. As we are interested in maintaining a minimum desired level of recall ℓ, we also consider the proportion of topics in the run where this minimum recall level is met; if this proportion is greater than or equal to a preset desired proportion p, we say the method is **reliable**. Where appropriate for the probabilistic methods, p will be viewed as the method's probability parameter.

5 RESULTS

Results are shown in Table 1.[4] All methods were reliable for all recall levels at both values of p. Mean recall levels were generally high for all methods. Mean recall for the knee method is always 1 since its parameters are set to the default levels, rather than being adjusted for particular values of p and ℓ. Although mean recall is reduced for other approaches, it always exceeds the desired recall levels (ℓ), particularly for the approaches based on counting processes (IP-E, IP-P, CX-E and CX-P).

The stopping criteria based on counting processes perform well in terms of effort / percentage of effort saved, with one of these methods producing the best results for all recall levels with $p = 0.95$ and the highest recall levels for $p = 0.8$. Performance of these methods is much better when the power law is used as the rate function and the improvement over the exponential rate function becomes more pronounced for higher levels of recall and confidence. For example, at the highest levels of recall and confidence using the power law rather than the exponential rate function with the Cox Process (CX-P and CX-E) reduces the number of documents that need to be examined by over 54 thousand. Using CX-P saves 46.1% of documents being examined while, in contrast, the target method saves 13.1% and IP-E only 0.9%.

Overall, there appears to be little difference between performance of the two counting processed compared (Inhomogenous Poisson Process, IP, and Cox Process, CX).

6 CONCLUSION

Determining when to stop making expensive manual relevance judgements is an important and challenging problem in Technology Assisted Review. Counting processes, such as Poisson and Cox processes, represent a theoretically motivated approach to developing stopping algorithms. This work demonstrated that modelling the rate at which relevant documents occur in a ranked set of retrieved documents using a power law is more effective than existing approaches which used exponential functions. Evaluation on the CLEF 2017 e-Health Lab Task 2 demonstrated that this approach worked well, particularly for high recall levels, and was able to achieve mean recall of 100% while avoiding the need to examine over 45% of the documents.

REFERENCES

[1] Gordon Cormack and Maura Grossman. 2016. Engineering Quality and Reliability in Technology-Assisted Review. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (Pisa, Italy). 75–84.

[4] A similar patterns of results was observed in further experiments carried out with additional levels of desired recall and confidence.

[2] Gordon V. Cormack and Maura R. Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. *CoRR* abs/1504.06868 (2015).

[3] Gordon V. Cormack and Maura R. Grossman. 2017. Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2017. *CoRR* abs/1504.06868.

[4] David R Cox. 1955. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)* 17, 2 (1955), 129–157.

[5] David R. Cox and Valerie Isham. 1980. *Point processes.* Vol. 12. CRC Press.

[6] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In *Working Notes of Clef 2017 - Conference and Labs of the Evaluation Forum*. Dublin, Ireland.

[7] John F C Kingman. 1993. *Poisson Processes.* Oxford : Clarendon Press.

[8] Dan Li and Evangelos Kanoulas. 2020. When to Stop Reviewing in Technology-Assisted Reviews: Sampling from an Adaptive Distribution to Estimate Residual Relevant Documents. *ACM Trans. on Information Systems* 38, 4 (2020), 1–36.

[9] David E. Losada, Javier Parapar, and Alvaro Barreiro. 2019. When to stop making relevance judgments? A study of stopping methods for building information retrieval test collections. *Journal of the Association for Information Science and Technology* 70, 1 (2019), 49–60.

[10] Alejandro Moreo and Fabrizio Sebastiani. 2019. Learning to Quantify: Estimating Class Prevalence via Supervised Learning. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France.

[11] Stephen E Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* (1977).

[12] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a “Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *Proceedings of the 31st International Conference on Distributed Computing Systems Workshops*. Washington, DC, USA, 166–171.

[13] Alison Sneyd and Mark Stevenson. 2019. Modelling Stopping Criteria for Search Results using Poisson Processes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 3482–3487.

[14] Byron C Wallace, Issa J Dahabreh, Kelly H Moran, Carla E Brodley, and Thomas A Trikalinos. 2013. Active literature discovery for scoping evidence reviews: How many needles are there. In *KDD workshop on data mining for healthcare*. 5–32.

[15] Haotian Zhang, Jimmy Lin, Gordon V Cormack, and Mark D Smucker. 2016. Sampling strategies and active learning for volume estimation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 981–984.

[16] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 307–314.