An Initial Analytical Exploration of Retrievability

Aldo Lipani[1] Mihai Lupu[2] Akiko Aizawa[1] Allan Hanbury[2]

[1]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan
{surname}@nii.ac.jp

[2]Inst. of Software Technology & Interactive Systems
Vienna University of Technology
Vienna, Austria
{surname}@ifs.tuwien.ac.at

ABSTRACT

We approach the problem of retrievability from an analytical perspective, starting with modeling conjunctive and disjunctive queries in a boolean model. We show that this represents an upper bound on retrievability for all other best match algorithms. We follow this with an observation of imbalance in the distribution of retrievability, using the Gini coefficient. Simulation-based experiments show the behavior of the Gini coefficient for retrievability under different types and lengths of queries, as well as different assumptions about the document length distribution in a collection.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*

Keywords

accessibility, retrievability, boolean model, Gini coefficient

1. INTRODUCTION

In the recent years a number of works have proposed and advocated to evaluate Information Retrieval (IR) models not just from the point of view of their efficacy but also from the accessibility that the model gives to the documents [1]. All these studies are fundamentally empirical, and no theoretical analysis has been done yet.

Accessibility plays a particularly important role in recall oriented domains. For example, patent experts are concerned about the fact that certain IR systems are biased towards particular patents rather than others. Also in the medical domain, medical researchers, while doing systematic reviews, to avoid such a bias include in their protocol the use of different search engines.

In essence, a retrievability study consists in automatically generating a number of queries, issuing them to an IR system, then counting how many times a document has been retrieved. Each step of the process has different parameters useful to characterize the IR system: the likelihood that a query, the parameters of the IR model, and at which rank a document is considered retrievable.

The idea of this paper comes from the recent experimental discoveries in which it has been pointed out that given an IR model, the length of the document influences its accessibility [5]. In the present study we therefore explore, under some assumptions, to which degree this happens. We do so analytically and through simulations. We start with the analysis of the perfect match models (boolean models), never conducted in previous retrievability studies. We then bridge the discoveries to the best match models, thanks to a small theoretical result that states their relationship.

The remainder of the paper is structured as follows: Section 2 provides the intuition of our method and introduces the required concepts. And we plot and briefly discuss the results in Section 3.

2. ANALYSIS

The retrievability concept is summarized by a measure [2] that defines how likely it is that a document is retrieved. Formally, the retrievability r of a document d with respect to a set of queries Q filed on a particular IR system, is defined as:

$$r(d) = \sum_{q \in Q} o_q f(d, q, c) \tag{1}$$

where o_q is the opportunity of the query being chosen, q a query, and f a utility function that measures how retrievable the document d is for a query q given the rank cut-off c. It is common to use as utility function f a function that gives 1 if the document is retrieved with rank above or equal to the cut-off c, and 0 if below.

For this initial study we focus however on the boolean search model. In this context, the outcome of the system is not a ranked list of documents but rather a set.

In previous retrievability studies, the queries Q have been generated following one of two strategies: a) starting from the indexed terms, for single term queries, all the terms that appear in the collection at least 5 times; for bi-terms queries, each bi-gram in the collection [2] b) starting from the documents, extracting all the bi-grams from the collection, and selecting those that appear more than 20 times [3, 4]. In both cases, the adopted procedure is an approximation of the entire set of possible queries.

ICTIR'15, September 27 - 30, 2015, Northampton, MA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3833-2/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808194.2809495 .

The study of boolean models does not require the generation of all the possible queries. We only need some assumption the class of query used. There are only a few characteristics of a query: it is length in terms, whether it is a unigram or n-gram, and whether it is conjunctive or disjunctive. Therefore, given the type of queries, we set off to analytically calculate the expected $r(d)$ for each document that has a specific number of unique terms.

2.1 Retrievability

Before going further, we introduce the notation used: T_d refers to the set of unique terms in the document d, T_q is the set of unique terms in the query q, T_c is the set of unique terms in the collection (i.e. the dictionary of the collection), and N_c the number of documents in the collection. $\sigma(d,q) : D \times Q \to \mathbb{R}$ is a scoring function assigning higher scores to more relevant documents, and

$$rank(d,q) = \begin{cases} |\{x \in D, \sigma(x,q) > \sigma(d,q)\}|, & \text{if } \sigma(d,q) > 0. \\ \infty, & \text{otherwise.} \end{cases} \tag{2}$$

a ranking function.

A Boolean model (also referred to sometimes as a perfect match model) is defined in the usual way: it considers relevant (and returns) a document matching the (sub)set of terms in the query. A best match model is essentially a ranking model applied on top of a Boolean model. Therefore, in this study we do not consider those ranking models which bypass individual terms and do their similarity computation in an abstract semantic space (e.g. Latent Semantic Indexing and Latent Dirichlet Allocation). In other words, a best match model here is any model where the implementation can be done using an inverted list and a weighting method.

2.1.1 The Conjunctive Case

For conjunctive queries all the query terms are required in order to retrieve a specific document. Given n, the size of the query, we can calculate $r(d)$ by interpreting the components of Eq. 1. The opportunity to use query q, o_q, is generally defined as Eq. 1 in Azzopardi's and colleagues' work [2, 6]. In this case, we can focus on the function f. We shall come back to o_q shortly.

The utility function f is essentially an indicator function with codomain in $\{0,1\}$ if its parameter is false or true. For a Boolean model, the utility function is therefore $f_B(d,q,c) = I(T_q \subseteq T_d)$, while for a best match (ranking) model, it is $f_R(d,q,c) = I(rank(d,q) < c)$.

For a random document d and query q, in the case of the Boolean model, the expectation of the utility function is the probability $P(T_q \subseteq T_d)$, which can be calculated by considering all possible sets of n terms ($|T_q| = n$) from the collection dictionary:

$$P(T_q \subseteq T_d) = \binom{|T_d|}{n}\binom{|T_c|}{n}^{-1}$$

Therefore, in the case of $o_q = 1$:

$$r(d) = \sum_{q \in Q_n} \binom{|T_d|}{n}\binom{|T_c|}{n}^{-1}$$

Given that $|Q_n| = \binom{|T_c|}{n}$, we finally have:

$$r(d) = \binom{|T_d|}{n} \tag{3}$$

However, if o_q was considered 1 for practical reasons in simulations, in this theoretical exercise where we already assumed that the vocabulary is limited by the collection vocabulary, we can estimate the probability of a query of length n as $\binom{|T_c|}{n}^{-1}$. Feeding that in the equation above, we obtain:

$$r(d) = \sum_{q \in Q_n} \binom{|T_c|}{n}^{-1}\binom{|T_d|}{n}\binom{|T_c|}{n}^{-1}$$

and following the same motivation as above:

$$r(d) = \binom{|T_d|}{n}\binom{|T_c|}{n}^{-1} \tag{4}$$

This is closer to a probabilistic perspective of retrievability, but in what follows we shall continue to use the form of Eq. 3 because, on one hand, it is simpler, and on the other hand, it is closer to what related empirical studies have been working with.

Now, let us consider all possible query sizes, that is with n that goes from 1 to $|T_c|$ the size of the test collection. The retrievability of a document d in case of using any combinations of n terms as conjunctive queries, is zero if $n > |T_d|$. Otherwise:

$$r(d) = \sum_{n=1}^{|T_c|}\binom{|T_d|}{n} = \sum_{n=1}^{|T_d|}\binom{|T_d|}{n} = 2^{|T_d|} - 1$$

2.1.2 The Disjunctive Case

When the queries are filed in disjunction, this means that at least one query term is required to retrieve a document. Given n, the size of query, we can calculate $r(d)$ similarly to the conjunctive case above. In this case, the utility function is $f_B(d,q,c) = I(T_q \cap T_d \neq \emptyset)$. Again, the expectation of this function is given by the probability

$$P(T_q \cap T_d \neq \emptyset) = \sum_{i=1}^{min(|T_q|,|T_d|)} P(|T_q \cap T_d| = i)$$

and consequently for any query of length $n \leq min\,(|T_q|, |T_d|)$, we have:

$$r(d) = \sum_{i=1}^{n}\binom{|T_d|}{i} \tag{5}$$

When considering queries of different lengths:

$$r(d) = \sum_{n \in N}\sum_{i=1}^{n}\binom{|T_d|}{i} \tag{6}$$

And, when considering all possible query lengths we have:

$$r(d) = \sum_{n=1}^{|T_d|}\sum_{i=1}^{n}\binom{|T_d|}{i}$$

2.1.3 Best-match models

Now, moving on to best match models, it becomes difficult to analytically consider their retrievability. However, a first observation is given in the theorem and corollary below.

THEOREM 1. *The retrievability of a document under a Boolean retrieval model B is an upper bound for the retrievability of the same document and the same query types, under any ranking system R.*

PROOF. $r_R(d) \leq r_B(d) \Leftrightarrow$
$\sum_{q \in Q} o_q f_R(d,q,c) \leq \sum_{q \in Q} o_q f_B(d,q,c) \Leftrightarrow$
$f_R(d,q,c) \leq f_B(d,q,c)$

For the conjunctive case, we have that the above is equivalent to $I(rank(d,q) < c) \leq I(T_q \subseteq T_d)$

Now, assume the contrary, $I(rank(d,q) < c) > I(T_q \subseteq T_d)$.

$\Leftrightarrow I(rank(d,q) < c) = 1$ and $I(T_q \subseteq T_d) = 0$.

similarly, for the disjunctive case we would have

$\Leftrightarrow I(rank(d,q) < c) = 1$ and $I(T_q \cap T_d \neq \emptyset) = 0$.

Both contradict our definition of the ranking function in Eq. 2 □

COROLLARY 1. *When there is no cutoff ($c = |N_c|$), the retrievability of a document in any of the best match models is equal to its retrievability in the Boolean model.*

2.1.4 N-Grams

In the analysis so far we have only considered queries of various sizes, but not with multi-word terms (n-grams). However, since n-grams are essentially terms in themselves, the only thing that would change is the scale of the calculation, rather than the observations about the nature of retrievability itself. We would agree that a more in-depth study into retrievability with n-grams is desirable, if only to prove our statement above, but we do make this simplification for this particular study.

2.2 Gini Coefficient

The purpose of this section is to observe the distribution of retrievability not over documents but rather over document lengths, counted in unique terms. This is because we want to observe the effect of the latter on the former, but also because in the current analytical view, two documents with the same number of unique terms are indistinguishable. To assess the bias of an IR model it is possible to observe the Lorenz curve, which visualizes the inequality among documents within a collection. The Lorenz curve has already been introduced in retrievability studies as the cumulative distribution of $r(d)$ ordered in non-decreasing order with varying of d. The Gini coefficient was proposed as a way to summarize with a single value the amount shown by the Lorenz curve [2, 6]. It is defined as:

$$G = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n}(n+1-i)y_i}{n\sum_{i=1}^{n} y_i} \tag{7}$$

where y_i is the population indexed in non-decreasing order ($y_i \leq y_{i+1}$), and n is the size of the population.

As we have observed in the previous analysis, $r(d)$ for a perfect match model is a function of the number of unique terms in the document. It can be in fact shown that $r(d)$ is monotonically increasing with $|T_d|$. Therefore, given a distribution of document lengths (based on unique terms) in a collection of documents, with probability mass function $u(s) = P(S = s)$, where S is the length of a document counted in unique terms, and $n = N_c$, the numerator in Eq. 7 is:

$$\sum_{i=1}^{N_c}(N_c{+}1{-}i)r(d_i) = \sum_{i=1}^{\infty}\sum_{j=1}^{\phi(i)}\left[N_c{+}1{-}\left(j{+}\sum_{s=1}^{i-1}\phi(s)\right)\right]r(d_{\phi(i)})$$

where $\phi(i) = \lfloor N_c u(i) + 1/2 \rfloor$ is the expected number of documents of length i, and $d_{\phi(i)}$ is a document of length i. The denominator is substituted by:

$$\sum_{i=1}^{N_c} r(d_i) = \sum_{i=1}^{\infty}\sum_{j=1}^{\phi(i)} r(d_{\phi(i)})$$

Simplifying, we obtain:

$$G = \frac{N_c+1}{N_c} - \frac{2\sum\limits_{i=1}^{\infty}[N_c + \frac{1}{2} - (\sum\limits_{s=1}^{i-1}\phi(s) + \frac{\phi(i)}{2})]\phi(i)r(d_{\phi(i)})}{N_c\sum\limits_{i=1}^{\infty}\phi(i)r(d_{\phi(i)})}$$

3. DISCUSSION AND CONCLUSION

With this definition of the Gini coefficient (G), we can now observe the effects of the query length and type (via $r(d)$) and of the distribution of document lengths in the collection (via $u(s)$). We do this in Fig. 1. We observe that the inequality always increases with the query length, the slope depends on the distribution of document lengths, and that the query type has a negligible effect. This last observation is potentially surprising. We explore this in Fig. 2. Here, the left-most element shows that the results become more skewed in different ways: in the conjunctive case a majority of documents get $r(d)$ equals to zero, while in the disjunctive case, a majority obtain high scores, when varying the length of the document. The other two plots show $r(d)$ for various query lengths n. In the case of single term queries, retrievability is essentially document length (again, counted as number of unique terms).

We have shown that retrievability for the Boolean model can be approached analytically. While in this study we considered different probability distributions for document lengths, the method can also be used in the presence of an actual test collection to calculate accessibility without the need for generating large sets of synthetic queries. Furthermore, the relationship between document length and retrievability, even in this particular retrieval model, may provide insights into new normalization factors for best match models.

4. REFERENCES

[1] L. Azzopardi and V. Vinay. *Accessibility in Information Retrieval*, volume 4956. Springer, 2008.

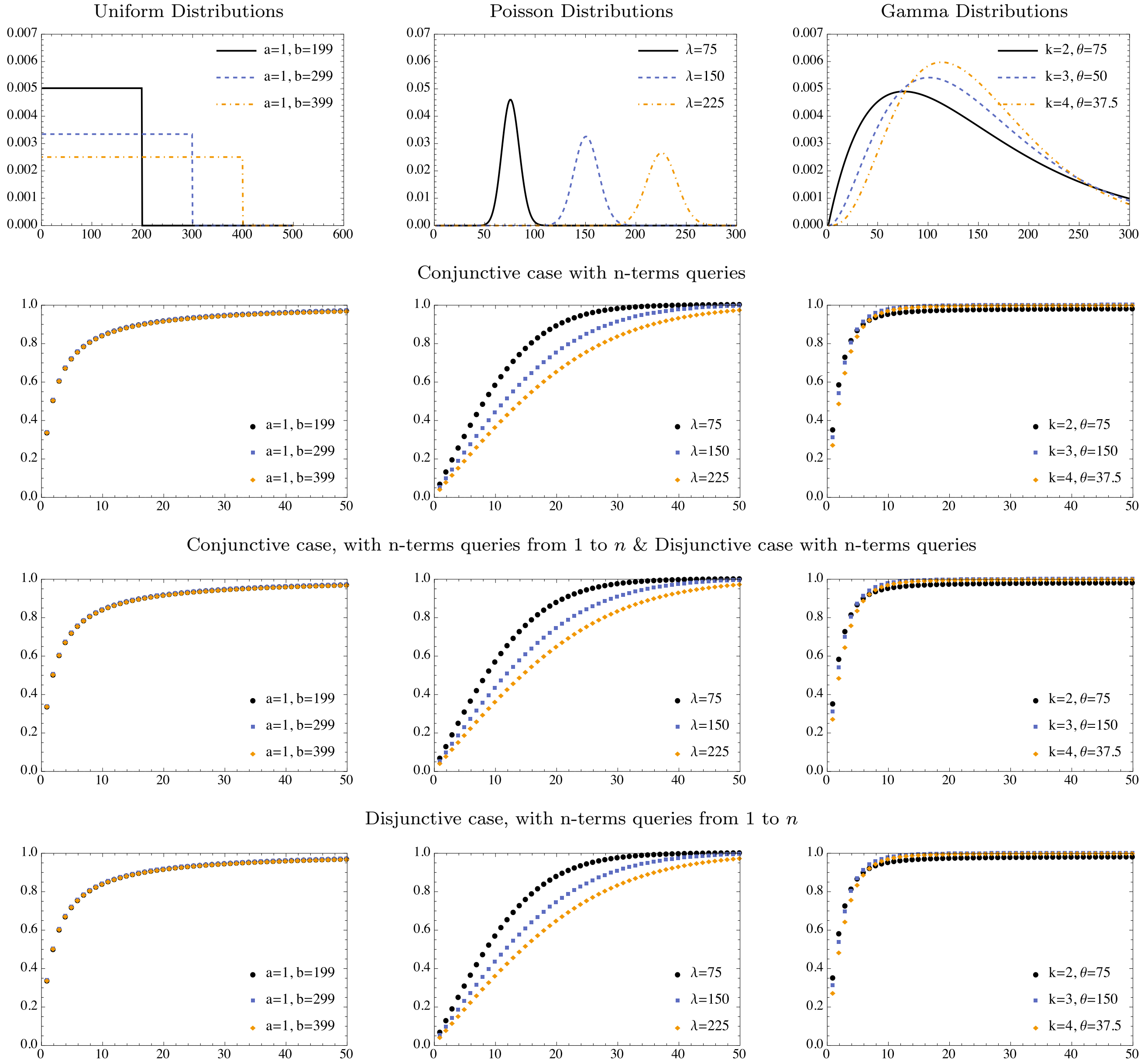
[2] L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proc. of CIKM*. ACM, 2008.

[3] S. Bashir and A. Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proc. of CIKM*, New York, NY, USA, 2009. ACM.

[4] S. Bashir and A. Rauber. *Improving Retrievability of Patents in Prior-Art Search*, volume 5993. Springer, 2010.

[5] C. Wilkie and L. Azzopardi. Relating retrievability, performance and length. In *Proc. of SIGIR*. ACM, 2013.

[6] C. Wilkie and L. Azzopardi. *Retrievability and Retrieval Bias: A Comparison of Inequality Measures*, volume 9022. Springer, 2015.


Uniform Distributions

Poisson Distributions

Gamma Distributions

Conjunctive case with n-terms queries

Conjunctive case, with n-terms queries from 1 to n & Disjunctive case with n-terms queries

Disjunctive case, with n-terms queries from 1 to n

Figure 1: Gini coefficient, given a document length distribution of a collection of documents vs different cases with varying of n of n-term queries. The middle row shows the two cases, conjunctive case, with n-terms queries from 1 to n and disjunctive case with n-terms queries, based on the observation that Eq. 5 models them both.

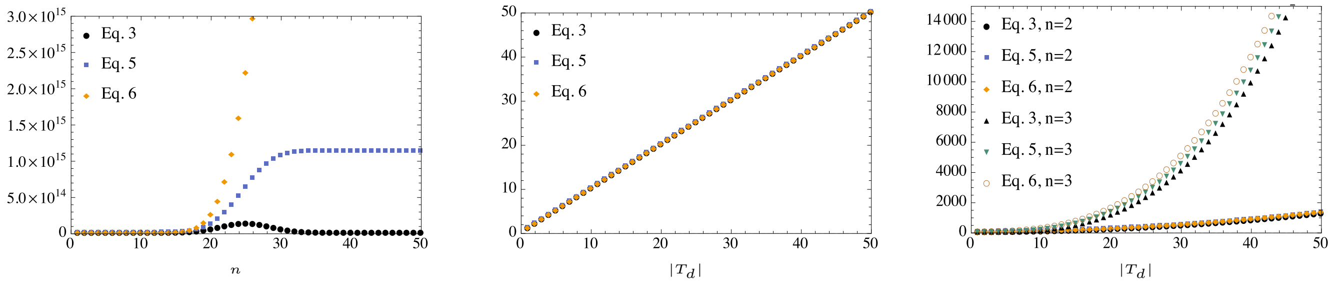


Figure 2: Retrievability $r(d)$ in the three cases: conjunctive with varying of n (Eq. 3), conjunctive with varying of n from 1 to n & disjunctive with varying of n (Eq. 5) and disjunctive from with varying of n from 1 to n (Eq. 6). The first plot to the left shows how the retrievability varies for a document of $|T_d| = 50$ with varying of n query terms; The second plot shows how the retrievability varies with varying of $|T_d|$ for single term queries; The third plot is similar to the second plot but with n equal to 2 and 3.