Information Processing and Management 48 (2012) 1171–1180

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

The tipping point: *F*-score as a function of the number of retrieved items

Raf Guns [a,*], Christina Lioma [b], Birger Larsen [c]

[a] University of Antwerp, IBW, Venusstraat 35, B-2000 Antwerpen, Belgium
[b] University of Copenhagen, Department of Computer Science, Njalsgade 128, DK-2300 Copenhagen, Denmark
[c] Royal School of Library and Information Science, Birketinget 6, DK-2300 Copenhagen S, Denmark

ARTICLE INFO

Article history:
Received 24 June 2011
Received in revised form 21 February 2012
Accepted 26 February 2012
Available online 19 March 2012

Keywords:
F-score
Precision
Recall
Information retrieval evaluation

ABSTRACT

One of the best known measures of information retrieval (IR) performance is the *F*-score, the harmonic mean of precision and recall. In this article we show that the curve of the *F*-score as a function of the number of retrieved items is always of the same shape: a fast concave increase to a maximum, followed by a slow decrease. In other words, there exists a single maximum, referred to as the tipping point, where the retrieval situation is 'ideal' in terms of the *F*-score. The tipping point thus indicates the optimal number of items to be retrieved, with more or less items resulting in a lower *F*-score. This empirical result is found in IR and link prediction experiments and can be partially explained theoretically, expanding on earlier results by Egghe. We discuss the implications and argue that, when comparing *F*-scores, one should compare the *F*-score curves' tipping points.

1. Introduction

In information retrieval (IR), one generally distinguishes between retrieved and non-retrieved items on the one hand and between relevant and non-relevant items on the other. Very generally, the goal of any IR method is to retrieve only relevant items – often, but not always, as many as possible. There exists a large collection of measures to evaluate the performance of a given method. In the context of this article, we assume binary relevance, where an item is either relevant or not relevant.

The two best known IR evaluation measures are precision and recall. Precision is defined as

$$P = \frac{|\text{ret} \cap \text{rel}|}{|\text{ret}|} \tag{1}$$

where ret denotes the set of retrieved items and rel denotes the set of relevant items. Recall is defined as

$$R = \frac{|\text{ret} \cap \text{rel}|}{|\text{rel}|} \tag{2}$$

Precision and recall are generally regarded as complementary (Buckland & Gey, 1994). In other words, one needs both to adequately assess the performance of a retrieval system. The reasons are well-known: it is, for instance, easy to achieve maximum recall by simply retrieving all items available in the system.

Several derived measures have been proposed that summarize precision and recall into one single number. Here, we study one of them, the so-called *F*-measure or *F*-score. The *F*-score is defined as the harmonic mean of precision and recall:

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R} \tag{3}$$

* Corresponding author.
E-mail addresses: raf.guns@ua.ac.be (R. Guns), liomca@gmail.com (C. Lioma), blar@iva.dk (B. Larsen).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2012.02.009

It is still one of the main performance measures in modern information retrieval (Manning, Raghavan, & Schütze, 2008), for instance in the legal IR domain (Hedin, Tomlinson, Baron, & Oard 2009). The *F*-score has the attractive property that if either precision or recall is low, *F* will be low as well; this is not the case when using the arithmetic or geometric mean. To the best of our knowledge, the foundation for the *F*-score was laid by van Rijsbergen (1974) and van Rijsbergen (1979), who introduced an IR effectiveness measure *E* (p. 128). The measure *E* also occurs in (Salton & McGill, 1983, p. 180). In essence, it is simply the complement of the *F*-score: $E = 1 - F$. Note that the *F*-score is equal to the Dice coefficient of the sets ret and rel:

$$F = 2\frac{|\text{ret} \cap \text{rel}|}{|\text{ret}| + |\text{rel}|} \tag{4}$$

A generalized (weighted) variant of *F* allows for assigning different weight to precision or recall:

$$F_\beta = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \text{ where } \beta^2 = \frac{1-\alpha}{\alpha} \tag{5}$$

where $0 \leqslant \alpha \leqslant 1$. Hence, β is a positive number. If $\beta > 1$, precision is emphasized over recall. If $\beta < 1$, recall is emphasized over precision. Eq. (3) uses F_1 with $\beta = 1$ or $\alpha = 1/2$, such that precision and recall have the same weight.

Many retrieval systems order results in terms of decreasing probability of relevance to a query. That is, items are assigned a relevance score, such that the items considered most likely to be relevant are presented first to the user. One may thus construct a curve of any evaluation measure as a function of the number of retrieved items $t = |\text{ret}|$. In this article we construct $F(t)$ curves, the *F*-score as a function of the number of retrieved items, and we show empirically and theoretically that *F*-score curves have a remarkable shape: a sharp concave increase, followed by a longer – usually convex – decrease. The maximum *F*-score is thus reached where the increase becomes a decrease; we refer to this maximum as the *tipping point*. Thus, if one accepts the *F*-score as a valid retrieval evaluation measure, the optimal situation occurs at the tipping point. The theoretical results offer a partial explanation of the empirically found regularities.

We briefly recall the definitions of concavity and convexity. A continuous function $f(x)$ is strictly concave on an interval if $f(\frac{x+y}{2}) > \frac{f(x)+f(y)}{2}$ for any x and y in the interval. It is strictly convex on an interval if $f(\frac{x+y}{2}) < \frac{f(x)+f(y)}{2}$ for any x and y in the interval. Informally, the curve is concave if it is always above a straight line between two points x and y in the interval, and convex if it is always below.

The remainder of the article is structured as follows. In the next section, we demonstrate $F(t)$ curves empirically found in IR and link prediction experiments. In Section 3, the shape of the *F*-score curves is partially explained using a theoretical model introduced by Egghe (2008); we will discuss the shape of the *F*-score curve in the cases of 'perverse', 'perfect', 'random' and 'normal' retrieval. Section 4 contains a further discussion of the implications of these findings. Finally, in Section 5 we present the conclusions.

2. Empirical results

Measures like recall and precision are defined on the basis of a 'contingency table', divided into four cells (Table 1). Since the same table can also be used for other applications, it is possible to apply *R* and *P* (and hence *F*) outside information retrieval. The empirical results presented in this section have been obtained from an IR application and a link prediction application, where we first encountered the regularities described in this paper. Similar results would likely have been obtained from other applications of these measures. In other words, although the paper mainly uses IR terminology, one should keep in mind that these measures can be used outside IR as well.

2.1. Information retrieval experiment

Given a user information need ('query'), the task of an IR system is to retrieve documents that are relevant to the query from a usually large repository of unstructured and heterogeneous documents, such as the Web. Our retrieval experiment uses a total of 100 queries from the Text Retrieval Evaluation Conference (TREC) Web track (queries 1–50 from 2009 and 51–100 from 2010) (Clarke, Craswell, & Soboroff 2009; Clarke, Craswell, Soboroff, & Cormack 2010) and the ClueWeb09 category B data set, which contains approximately 50 million web pages in English crawled between January and February 2009. Indexing and retrieval is realized with the Indri open source IR system (Strohman, Metzler, Turtle, & Croft, 2005). We use three different retrieval models from the language modeling (LM) framework (Croft & Lafferty, 2003): (i) LM with Dirichlet

Table 1
Contingency table for information retrieval and related applications.

	Retrieved	Not retrieved
Relevant	True positive	False negative
Not relevant	False positive	True negative

smoothing, (ii) LM with Jelinek–Mercer smoothing, and (iii) LM with two-stage Dirichlet and Jelinek–Mercer smoothing. Dirichlet includes a parameter μ, and Jelinek–Mercer includes a parameter λ. We tune these parameters using 5-fold validation separately for each query set on Zhai and Lafferty's (2002) tuning ranges: $\mu \in \{100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10{,}000\}$, $\lambda \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$. For all results, we record the number of relevant documents, as well as the number of retrieved items that are relevant at each step of 10 retrieved documents (up to 1500). This enables us to determine the evolution of F as a function of the number of retrieved items. Query number 95 was excluded from the subsequent analysis, because no documents were assessed relevant for this query. In the case of query number 40, only 128 documents were retrieved, because that query contains one very rare word ("michworks").

It is practically difficult to show resulting F-score curves for all 100 information needs in the three retrieval models. We therefore give an overview of the variety of results and some overall properties. Fig. 1 shows the curve of the F-score as a function of the number of retrieved items for five queries (Q19, Q35, Q46, Q74, and Q86) using LM with Dirichlet smoothing. The five queries have been selected to represent the broad variety of resulting curves. Q35 and Q46 first exhibit a concave increase to a maximum, followed by a largely convex decrease. Q35 is slightly less regular than Q46, in that the first part of the curve has some occasional local decreases. Q86 follows a similar pattern, although it is very irregular at the top, with two clear local maxima. Q19's curve is again similar, although the decreasing part seems to remain virtually horizontal. Nevertheless, examination at a local scale reveals that the maximum F-score in fact occurs in the left-most part of the figure. Finally, Q74 does not really obey the expected pattern (according to the theoretical analysis, Section 3), mostly because initially the F-score remains at zero. Curves like the one for Q74 occur mainly in those cases where the number of relevant documents is very small. The tipping points in Fig. 1 are reached at different numbers of retrieved documents, which is due to the fact that each curve corresponds to a separate query.

We do not show the curve of the F-score as a function of the number of retrieved items for the above five queries using LM with two-stage smoothing, because it is very similar to Fig. 1. However, Jelinek–Mercer smoothing yields curves that overall exhibit less irregularities than the other two models, as illustrated in Fig. 2. This can be seen most clearly by comparing Q86 in the two Figures. However, in most cases the maximum F-score is lower for Jelinek–Mercer smoothing than for the other two models.

These curves illustrate that, in general, the curve of the F-score has a distinct shape, which will be theoretically explained in Section 3. However, it can be seen that curves of individual queries may deviate from the expected shape. This is mostly due to fluctuations in precision, which in real-world cases is rarely a strictly decreasing function of the number of retrieved items – although ideally it is.

The experiment also reveals that maximum F-scores may vary a lot across different queries. Fig. 3 illustrates the variety across all queries. Since the results are approximately normally distributed, the mean or median of maximum F-scores can be used as a summary statistic of the IR system's overall performance, similarly to the way the average R-Precision of a TREC run is the mean of R-Precisions for each individual query in the run.

Another approach would be to create an aggregated F-score curve by taking the average over all queries for each number of retrieved items, as displayed in Fig. 4. This has the advantage of resulting in fairly regular curves with the same features. It is, however, unclear whether this is indeed a viable and methodologically sound approach.

2.2. Link prediction experiment

We briefly introduce the link prediction application. Essentially, link prediction tries to predict the future state of an evolving network based on an earlier snapshot (Liben-Nowell & Kleinberg, 2007). The earlier snapshot is known as the training network, which is the input to the prediction process. The prediction constructed on the basis of the training network is simply known as the predicted network. To assess the quality of the prediction, one can compare the predicted network with an actual later snapshot, called the test network. Predicted and unpredicted links are analogous to retrieved and non-retrieved items respectively; links present and not present in the test network are analogous to relevant and non-relevant

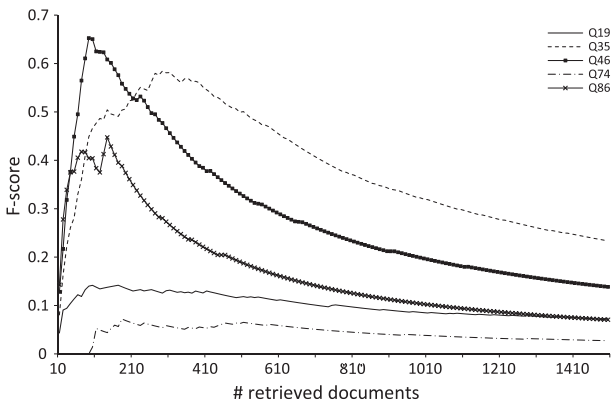

Fig. 1. F-score as a function of number of retrieved documents for five queries (LM with Dirichlet smoothing).

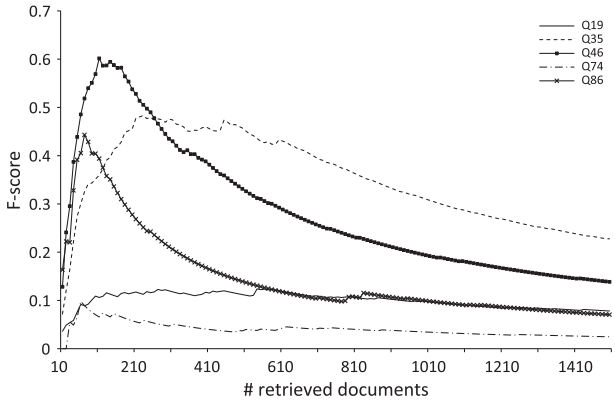

Fig. 2. *F*-score as a function of number of retrieved documents for five queries (LM with Jelinek–Mercer smoothing).

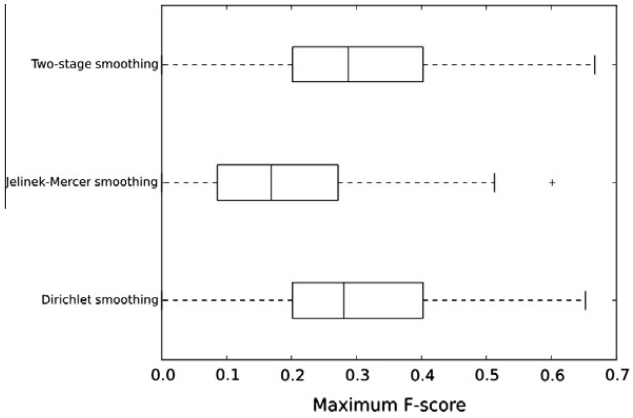

Fig. 3. Box plots of maximum *F*-scores for all queries for three LMs.

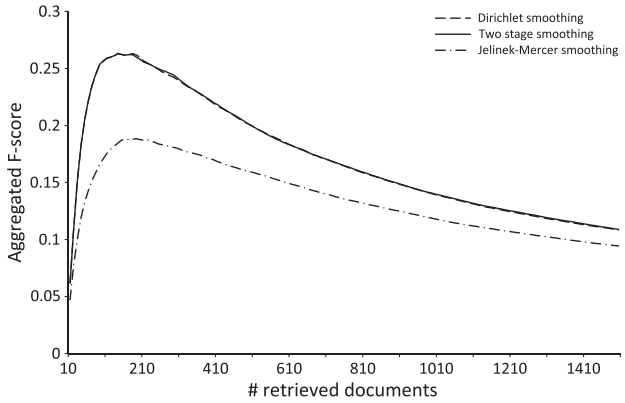

Fig. 4. *F*-score curves aggregated over all queries for three LMs.

items respectively. Because some links are more likely than others, we can order them in decreasing order of likelihood, similar to relevance ranking. Taken together, these elements imply that link prediction can be evaluated using IR evaluation measures, such as precision and recall.

Fig. 5 shows the *F*-score as a function of the number of predicted links for three predictors, applied to the 'AcadBib' data set. For more details about this data set and the predictors used we refer to (Guns, 2009; Guns, 2011; Liben-Nowell & Kleinberg, 2007). All three cases clearly show a steep concave increase to a maximum, followed by a slower decrease. The decrease is completely convex for the Graph distance and Katz predictors; Simrank's decreasing part, however, appears to start out concave and then turn into convex. Overall, the curves for link prediction are more regular than those for IR for this data set, an observation partially due to the different characteristics of the ClueWeb09 and AcadBib data sets.

Interestingly, each predictor reaches the tipping point around the same number of predictions. In this particular case, there can be little doubt that Simrank is outperformed by the other two, since the former's curve is below the others'

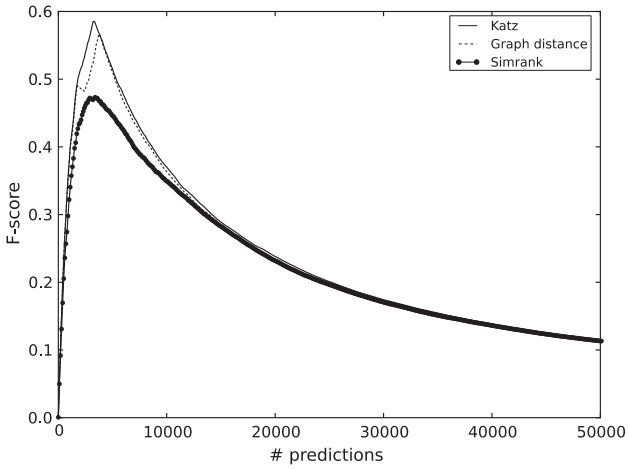

Fig. 5. *F*-score as a function of number of predicted links for three predictors.

throughout the chart. Graph distance is generally very similar to Katz, but reaches a slightly lower tipping point. Note that the differences between predictors are largest at their tipping points; further on, we will offer a simple explanation for this observation.

Similar results can be obtained using other predictors and data sets. We point out that Fig. 5 is truncated on the right, such that the environment around the tipping point is more clearly visible; the same convexly decreasing trend actually continues up to about 600,000 predictions. Some predictors – such as those that are based on common neighboring nodes – yield only a limited number of predictions. In those cases, it is possible that the curve ends before the tipping point is reached.

3. Theory: the *F*-score curve in different retrieval situations

In this section, we address the question why specific shapes such as those found in Fig. 1, 4 and 5 occur.

Egghe (2008) presents curves of the measures precision, recall, fallout and miss for some generalized retrieval situations, which were originally introduced by Buckland and Gey (1994). One can distinguish between *perverse retrieval* (first return all non-relevant items, then the relevant ones), *perfect retrieval* (first return all relevant items, then all non-relevant ones), *random retrieval* (randomly return items without regard for their relevance), and *normal retrieval* (the situation where the density of relevant items is higher for small t than for high t). In other words, normal retrieval models the usual situation where one tries to return only relevant items but inevitably makes some errors. Perverse, perfect and random retrieval are not very realistic, but they are theoretically valuable because they represent extreme cases of IR.

We will build upon Egghe's results to describe the *F*-score curve for these retrieval situations. First, we briefly summarize the theoretical framework we are working in (see Egghe (2008) for full details). We have $t = |\text{ret}|$ retrieved items. We assume a given query or IR problem, such that the number of relevant items, denoted as $\ell = |\text{rel}|$, is fixed. The entire system has N items. Now, for every t, the number of relevant items found so far can be counted. This number is denoted as $H(t)$. Contrary to the discrete reality (where $t = 1, 2, \ldots, N$), we use a continuous model, such that $t \in [0, N]$. Furthermore, we assume that the function H is differentiable and denote $H' = h$. It follows that

$$H(t) = \int_0^t h(i)di \tag{6}$$

Here, h is the retrieval density function, specifying the density of relevant items at each number of retrieved items t. In the continuous setting of Egghe (2008), the *F*-score is then equal to[1]

$$F = F(t) = \frac{2}{\frac{1}{P(t)} + \frac{1}{R(t)}} = \frac{2\int_0^t h(i)di}{t + \ell} \tag{7}$$

Note, again, the equivalence of the *F*-score (7) with the Dice coefficient (4).

Since $\int_0^0 h(i)di = 0$, in all cases, the curve starts at

$$F(0) = 0 \tag{8}$$

And since $\ell = \int_0^N h(i)di$, in all cases, the curve ends at

$$F(N) = \frac{2\ell}{N + \ell} \tag{9}$$

[1] Egghe (2008) uses *F* to denote fallout, another measure which is not used here.

3.1. Perverse retrieval

Perverse retrieval is the extreme case of the worst possible retrieval result: first, we find all non-relevant items and then the relevant ones.

Egghe (2008) has the following result:

$$\int_0^t h(i)di = t - \min(t, N - \ell) \tag{10}$$

The *F*-score function then is:

$$F(t) = 2\frac{t - \min(t, N - \ell)}{t + \ell} \tag{11}$$

Then, for $0 \leqslant t \leqslant N - \ell$:

$$F(t) = 0$$

And for $t > N - \ell$:

$$F(t) = 2\frac{t - N + \ell}{t + \ell}$$

The resulting curve is shown in Fig. 6. Where $t = N - \ell$, the curve is continuous, but not differentiable.

3.2. Perfect retrieval

In case of perfect retrieval, Egghe (2008) proves the following result:

$$\int_0^t h(i)di = \min(t, \ell) \tag{12}$$

Hence, the *F*-score is given by:

$$F(t) = 2\frac{\min(t, \ell)}{t + \ell} \tag{13}$$

Then, if $0 \leqslant t \leqslant \ell$:

$$F(t) = \frac{2t}{t + \ell}$$

And if $t > \ell$:

$$F(t) = \frac{2\ell}{t + \ell}$$

This yields a curve as shown in Fig. 6: we have a concave increase to $(\ell, 1)$, followed by a convex decrease to $\left(N, \frac{2\ell}{N+\ell}\right)$. Where $t = \ell$, the curve is continuous, but not differentiable.

3.3. Random retrieval

In case of random retrieval, we have that (Egghe, 2008):

$$\int_0^t h(i)di = \frac{\ell t}{N} \tag{14}$$

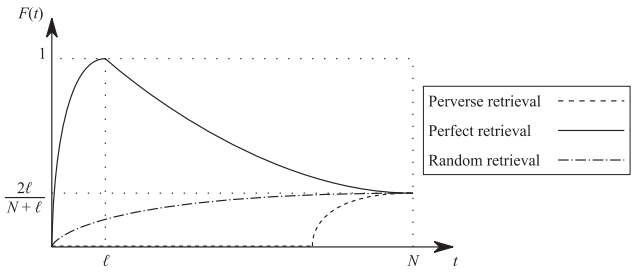

Fig. 6. $F(t)$ curves in case of perverse, perfect and random retrieval.

The *F*-score curve is then described by:

$$\begin{aligned} F(t) &= \frac{2\frac{\ell t}{N}}{t+\ell} \\ &= 2\frac{\ell t}{N(t+\ell)} \end{aligned} \tag{15}$$

This yields a curve as shown in Fig. 6.

3.4. Normal retrieval

Egghe (2008) defines normal retrieval as the retrieval situation where the following conditions hold.

$$h(0) = 1 \tag{16}$$

$$h(N) = 0 \tag{17}$$

Furthermore, the density of relevant items decreases as one retrieves more items:

$$h'(t) < 0 \tag{18}$$

Informally, Eq. (18) states that there are more relevant items among higher ranked items than among lower ranked ones. This is a very natural requirement for normal retrieval situations, related to van Rijsbergen's (1974) concept of decreasing marginal effectiveness. In theory it should not matter if one trades an increase in precision (or recall) for an equally sized decrease in recall (or precision). In reality however, there is no perfect trade-off: the marginal effectiveness of precision (or recall) is decreasing as one retrieves more items. Assumptions (16)–(18) can be interpreted as a consequence of the probability ranking principle (Robertson, 2007): an item with a higher score has a higher probability of relevance, where Eqs. (16) and (17) represent the limiting cases. Hence, since items are ranked in decreasing order of their score, the density of relevant items is a decreasing function of *t*.

To the three assumptions (16)–(18) we add the extra assumption that there is at least one relevant item in the system:

$$0 \leq \ell \leq N \tag{19}$$

This assumption is needed to avoid some cases of division by zero. Note that this assumption is also needed if one wants to avoid division by zero in the definition of recall. We remark that some of Egghe's calculations also rely on the assumption $h''(t) < 0$; since this assumption is not needed for our purposes, we do not use it here.

We now consider the question of the curve's shape. The first derivative is:

$$\begin{aligned} F'(t) &= \frac{2(\ell+t)h(t) - 2\int_0^t h(i)di}{(\ell+t)^2} \\ &= \frac{2}{\ell+t}\left(h(t) - \frac{\int_0^t h(i)di}{\ell+t}\right) \end{aligned} \tag{20}$$

The sign of $F'(t)$ is determined by the expression between brackets. We now examine the start and end of the curve.

As *t* approaches 0, $h(t)$ approaches 1. This follows from (16) and the continuity of *h*. Moreover, as *t* approaches 0, $\frac{\int_0^t h(i)di}{\ell+t}$ approaches 0 (assuming (19)). So for small *t*, we have

$$h(t) > \frac{\int_0^t h(i)di}{\ell+t} \tag{21}$$

Hence, $F'(t) > 0$ and $F(t)$ is increasing.

At the end of the curve, *t* approaches *N*. Here, $h(t)$ approaches 0 and $\frac{\int_0^t h(i)di}{\ell+t}$ approaches $\frac{\int_0^N h(i)di}{\ell+t} = \frac{\ell}{N+\ell}$. This is greater than 0, as follows from (19). For large *t*, we have

$$h(t) < \frac{\int_0^t h(i)di}{\ell+t} \tag{22}$$

Hence, $F'(t) < 0$ and $F(t)$ is decreasing.

Since $h(t)$ is first increasing (21) and then decreasing (22) and furthermore *h* is a continuous function, there exists a maximum point $t_0 \in]0, N[$ such that $F'(t_0) = 0$. In other words, in this continuous setting as well, we find the existence of a tipping point. Note, however, that we have not been able to show that there exists exactly one such maximum. It is theoretically possible that there is more than one maximum.

We now consider the second derivative of $F(t)$. From (20), it follows:

$$F''(t) = \left(\frac{1}{\ell+t}\right)\left(h'(t) - \frac{(\ell+t)h(t) - \int_0^t h(i)di}{(\ell+t)^2}\right) + \left(h(t) - \frac{\int_0^t h(i)di}{\ell+t}\right)\left(\frac{-1}{(\ell+t)^2}\right)$$
$$= \frac{h'(t)}{\ell+t} - \frac{h(t)}{(\ell+t)^2} + \frac{\int_0^t h(i)di}{(\ell+t)^3} - \frac{h(t)}{(\ell+t)^2} + \frac{\int_0^t h(i)di}{(\ell+t)^3} = \frac{h'(t)}{\ell+t} - 2\frac{h(t)}{(\ell+t)^2} + 2\frac{\int_0^t h(i)di}{(\ell+t)^3} \quad (23)$$

So the sign of $F''(t)$ is determined by the expression

$$h'(t) - 2\frac{h(t)}{\ell+t} + 2\frac{\int_0^t h(i)di}{(\ell+t)^2} \quad (24)$$

As t approaches 0, it follows from (16) and (19) that (24) approaches

$$h'(0) - 2\frac{h(0)}{\ell} + 2\times 0 = h'(0) - \frac{2}{\ell} < 0$$

Therefore, $F(t)$ always starts concavely.
For t approaching N, it follows from (17) that (24) approaches

$$h'(N) - 2\times 0 + 2\frac{\int_0^N h(i)di}{(\ell+N)^2} = h'(N) + 2\frac{\ell}{(\ell+N)^2} \quad (25)$$

One can see that $F(t)$ can end convexly or concavely. $F(t)$ ends convexly iff

$$h'(N) > -2\frac{\ell}{(\ell+N)^2} \quad (26)$$

and $F(t)$ ends concavely iff

$$h'(N) \leq -2\frac{\ell}{(\ell+N)^2} \quad (27)$$

4. Discussion

Given a set of reasonable assumptions, it can be shown that normal retrieval leads to an F-score curve that first increases concavely and then decreases, and has a maximum. However, some aspects of the empirically found curves are not explained by the model. These are: (1) in most cases, the curve appears to decrease convexly rather than concavely, (2) the decreasing part is much longer than the increasing part, and (3) for a given query (or test network) different IR strategies (or different predictors) reach a maximum around the same value of t (number of retrieved documents or predictions).

We can, however, suggest a heuristic explanation for these phenomena. Normal retrieval always tries to achieve perfect retrieval – one could state that perfect retrieval is a limiting case of normal retrieval. Hence, it is reasonable to assume that the results of normal retrieval will (to a certain extent) resemble the results of perfect retrieval. Indeed, the three unexplained phenomena can also be found in the case of perfect retrieval. First, the decreasing part of a perfect retrieval curve is always convex. Second, the tipping point of the perfect retrieval curve is reached when ℓ items have been retrieved. Since typically $\ell \ll N$, the increasing part is shorter than the decreasing one. Third, in case of perfect retrieval the tipping point is reached when $t = \ell$, where $P = R = 1$.[2] In case of normal retrieval, the tipping point is reached slightly after $t = \ell$ (judging by empirical data). Note that the other two extreme cases, random and perverse retrieval, reach a maximum only when $t = N$ – the fact that these two cases do not have a tipping point illustrates again that normal retrieval is qualitatively closer to perfect retrieval.

An advantage of the tipping point is that it is the only point where it is possible that $F(t) = 1$ (only in case of perfect retrieval, when $t = \ell$). Thus, it allows for a larger range than any other point and can better discriminate between different retrieval situations. This also explains our earlier observation that the difference between two F-score curves is largest at the tipping point. For this reason, and because the tipping point occurs around the same value of t (and close to ℓ) we suggest that, if one wants to compare F-scores in a ranked retrieval situation, one can best compare the respective tipping points.

When considering these findings from an IR point of view, a limitation of the current approach is that it assumes the existence of one individual query, rather than (as is customary in IR) aggregating over many queries. Moreover, it may be the case that observed rankings for individual queries are quite different from the continuous model – especially for those cases that have few relevant items, as discussed in Section 2.1. We first point out that the same limitations exist for Egghe's (2008) original work; the model is certainly a simplification but it has the advantage of highlighting some generally expected reg-

[2] Generally, the point where $P = R$ is known as the break-even point or R-Precision (Manning et al., 2008). Manning et al. mention that "it is somewhat unclear why you should be interested in the break-even point rather than either the best point on the curve (the point with maximal F-measure) or a retrieval level of interest to a particular application (Precision at k)". Our analysis suggests that the break-even point and the point with maximal F-measure are typically very close together.

ularities. In spite of its simplicity, such regularities can be found in empirical data, as shown in Section 2. We do not offer a complete solution, but possible approaches to these problems include determining the maximum *F*-score for each query and aggregating over that using a simple mean, analogous to mean average precision (MAP) or constructing an aggregated *F*-score curve and finding its maximum. More research would be necessary to determine the best approach.

There exists a significant amount of IR literature on modeling the distributions of scores of relevant and non-relevant documents (see Doloc-Mihu (2009) for an overview). The general motivation for this line of work is that the statistical properties of retrieval scores, displayed by the shape of their distribution, for a given query, can be used to model aspects of the retrieval process. From the early work of Swets (1963), who proposed to model score distributions to find an optimal threshold for separating relevant from non-relevant documents, to the more recent work of Dai, Kanoulas, Pavlu, and Aslam (2011), who used score distributions models for inferring precision-recall curves, various combinations of statistical distributions have been proposed; some focus more on their goodness of fit to some set of empirical data, while others consider more their theoretical properties (Robertson, 2007). Although closely related, this work differs in a few respects from research on score distributions. First, we do not take scores and their distributions directly into account, only the resultant ranking. Second, since we make only general assumptions and do not try to fit any particular score distribution, our theoretical part is rather general. The question of how specific score distributions affect the possible shapes of the *F*-score curve is interesting, but not studied here.

5. Conclusions

We have shown empirically that the *F*-score as a function of the number of retrieved items typically has a distinct shape with a clear maximum – the tipping point. Since the same shape occurs for different applications, data sets and IR strategies, this is clearly not a coincidence. The tipping point is the 'best' or 'optimal' result: retrieving either less or more items results in a lower *F*-score. We stress that terms like 'best' and 'optimal' should always be understood with respect to the *F*-score and that other IR performance measures may yield other results.

The empirical findings have spurred a theoretical investigation within the continuous framework of Egghe (2008). This has led to a characterization of the *F*-score curve in four different retrieval situations: perverse, random, perfect and normal retrieval. It can be shown that perfect and normal retrieval lead to the distinct shape with a tipping point, which was found empirically. Interestingly, the model seems to allow for a greater variety of shapes than found empirically. The empirical results are qualitatively closer to perfect retrieval than to random or perverse retrieval.

We close with some suggestions for future research. It would be interesting to examine *F*-score curves for more data sets to get a better view of possible variations that our experiments may have missed. Are there, for instance, empirical results that do yield a concavely ending *F*-score curve? Also, there exist other derived IR performance measures. What is their behavior as a function of the number of retrieved items? Finally, an important but difficult problem remains the question how a system should estimate the point at which *F* (or another performance measure) is maximized, if no relevance information is available.

Acknowledgements

We are grateful to Leo Egghe for his suggestions and his help regarding Section 3.4. Ronald Rousseau and two anonymous reviewers provided useful comments, which helped to improve the paper.

References

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science and Technology, 45*(1), 12–19.

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 web track. In E. M. Voorhees & L. P. Buckland (Eds.), *Proceedings of the eighteenth text retrieval conference, TREC 2009, Gaithersburg, Maryland, USA, November 17–20, 2009*. Gaithersburg: National Institute of Standards and Technology.

Clarke, C. L. A., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). *Overview of the TREC 2010 web track. Proceedings of the nineteenth text retrieval conference, TREC 2010, Gaithersburg, Maryland, USA, November 16–19, 2010*. Gaithersburg: National Institute of Standards and Technology.

Croft, W. B., & Lafferty, J. (2003). *Language modeling for information retrieval*. Kluwer Academic Publishers.

Dai, K., Kanoulas, E., Pavlu, V., & Aslam, J. A. (2011). Variational bayes for modeling score distributions. *Information Retrieval, 14*(1), 47–67.

Doloc-Mihu, A. (2009). Modeling score distributions. In J. Wang (Ed.) *Encyclopedia of data warehousing and mining*. Hershey, IGI Global (pp. 1330–1336).

Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing & Management, 44*(2), 856–876.

Guns, R. (2009). Generalizing link prediction: Collaboration at the University of Antwerp as a case study. *Proceedings of the American Society for Information Science & Technology, 46*(1), 1–15.

Guns, R. (2011). *Bipartite networks for link prediction: Can they improve prediction performance? Proceedings of ISSI 2011 – 13th international conference of the international society for scientometrics and informetrics*. Leiden: Leiden University.

Hedin, B., Tomlinson, S., Baron, J. R., & Oard, D. W. (2009). Overview of the TREC 2009 legal track. In E. M. Voorhees & L. P. Buckland (Eds.), *Proceedings of the eighteenth text retrieval conference, TREC 2009, Gaithesburg, Maryland, USA, November 17–20, 2009*. Gaithesburg: National Institute of Standards and Technology.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: University Press.

van Rijsbergen, C. J. (1974). Foundation of evaluation. *Journal of Documentation, 30*(4), 365–373.

van Rijsbergen, C. J. (1979). *Information retrieval* (second ed.). Dept. of Computer Science, University of Glasgow.

Robertson, S. (2007). On score distributions and relevance. In G. Amati, C. Carpineto, & G. Romano (Eds.), *Proceedings of the advances in information retrieval, 29th European conference on IR research, ECIR 2007, Rome, Italy, April 2–5, 2007* (pp. 40–51). Berlin: Springer-Verlag.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New york: McGraw-Hill.

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005). Indri: A language model-based search engine for complex queries. In *Proceedings of the international conference on intelligence analysis (ICIA 2005)*.

Swets, J. A. (1963). Information retrieval systems. *Science, 141*(3577), 245–250.

Zhai, C., & Lafferty, J. D. (2002). Two-stage language models for information retrieval. In M. Beaulieu, R. Baeza-Yates, & S. H. Myaeng (Eds.), *SIGIR 2002: Proceedings of the twenty-fifth annual international ACM SIGIR conference on research and development in information retrieval, August 11–15, 2002, Tampere, Finland* (pp. 49–56). New York: ACM Press.