

Приложение А.

Официальные метрики РОМИП 2006

М. Агеев, И. Кураленок, И. Некрестьянов
ageev@mail.cir.ru, ik@oasis.apmath.spbu.ru, nis@acm.org

Для оценки качества работы поисковой (рубрицирующей) системы применяются различные оценки, основанные на анализе результатов работы системы. При этом "идеальным" алгоритмом считается тот, для которого выводы, сделанные системой, согласуются с мнением оценивающих экспертов. В РОМИП'2006 использовались следующие метрики оценки качества работы систем:

- в дорожках поиска:
 1. полнота (recall)
 2. точность (precision)
 3. средняя точность (average precision)
 4. точность на уровне 5 документов (precision(5))
 5. точность на уровне 10 документов (precision(10))
 6. R-точность (R-precision)
 7. 11-точечный график полноты/точности, измеренный по методике TREC (11-point matrix (TREC))
 8. 11-точечный график полноты/точности, модифицированный вариант (11-point matrix (ROMIP))
- в дорожках классификации:
 1. полнота (recall)
 2. точность (precision)
 3. аккуратность (accuracy)
 4. ошибка (error)
 5. F-мера (F-measure)

- в дорожках аннотирования по запросу:
 1. точность (precision)
 2. аккуратность (accuracy)
 3. ошибка (error)
- в дорожке вопросно-ответного поиска:
 1. точность (precision)
 2. усредненная ценность ответов по линейке TREC (TrecReciprocalRank)
 3. усредненная ценность ответов по линейке РОМИП (RomipReciprocalRank)

Большая часть из этих метрик подробно изучена и описана в литературе [1, 3, 4, 5, 6], однако, интерпретации тех или иных оценок зачастую различаются. Поэтому мы подробно опишем каждую из этих метрик.

Сначала мы дадим описание метрик оценки качества работы системы в применении к одному запросу (рубрике), а затем опишем методики усреднения метрик для получения интегральных показателей качества поиска/классификации.

Для многих метрик спорным вопросом является случай, когда для данного запроса нет релевантных документов. Например, значение полноты в этом случае — неопределенность типа 0/0. В TREC такие запросы не учитываются при вычислении метрик [1]. В РОМИП принято такое же соглашение: запросы, для которых нет релевантных документов, не рассматриваются при вычислении метрик.

1. Метрики на множествах документов

Большинство метрик, применяемых в современной оценке текстового поиска, основываются на отношении релевантности (принадлежности) документа запросу (рубрике). Обсуждение самого понятия релевантности выходит за рамки данного документа. Здесь необходимо лишь отметить, что это отношение имеет скорее психологическую природу и устанавливается прямым опросом экспертов-оценщиков. Метрики для неупорядоченного множества документов основаны на бинарной классификации документов «релевантен/не релевантен» по отношению к выбранному запросу. Данные метрики основываются на матрице классификации:

	релевантны	не релевантны
найдено системой	a	b
не найдено системой	c	d

Таблица 1. Основные категории документов ответа системы

Здесь a — количество документов, найденных системой и релевантных с точки зрения экспертов; b — количество документов, найденных системой, но не релевантных с точки зрения экспертов; c — количество релевантных документов, не найденных системой; d — количество нерелевантных документов, не найденных системой.

1.1. Полнота (recall)

Полнота (recall) вычисляется как отношение найденных релевантных документов к общему количеству релевантных документов:

$$r = \frac{a}{a + c}$$

Полнота характеризует способность системы находить нужные пользователю документы, но не учитывает количество нерелевантных документов, выдаваемых пользователю. Например, если полнота равна 50%, то это значит, что половина релевантных документов системой не найдена.

1.2. Точность (precision)

Точность (precision) вычисляется как отношение найденных релевантных документов к общему количеству найденных документов:

$$p = \frac{a}{a + b}$$

Точность характеризует способность системы выдавать в списке результатов только релевантные документы. Например, если точность равна 50%, то это значит, что среди найденных документов половина релевантных и половина — нерелевантных.

1.3. Аккуратность (accuracy)

Аккуратность (accuracy) вычисляется как отношение правильно принятых системой решений к общему числу решений. Формально:

$$Accuracy = (a+d) / (a+b+c+d)$$

Поскольку мы предполагаем, что система принимает решение о принадлежности к данной категории для каждого документа коллекции. Таким образом знаменатель не зависит от рассматриваемой категории. При вычислении оценки в качестве знаменателя использовалось общее число документов оценивавшихся хотя бы для одной категории (что сильно меньше числа документов, но это сказывается лишь на масштабе).

Отметим, что микроусреднение и макроусреднение для этой оценки дают одинаковый результат.

1.4. Ошибка (error)

Ошибка (error) вычисляется как отношение неправильно принятых системой решений к общему числу решений. Формально:

$$Error = (b+c) / (a+b+c+d)$$

1.5. F-мера (F-measure)

F-мера часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера для данного запроса (рубрики) вычисляется по формуле

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Отметим основные свойства метрики F:

- $0 \leq F \leq 1$
- если $r = 0$ или $p = 0$, то $F = 0$
- если $r = p$, то $F = r = p$
- $\min(r, p) \leq F \leq \frac{r+p}{2}$

2. Усреднение множественных метрик

Одним из важных вопросов построения метрик в текстовом поиске (особенно в случае классификации и фильтрации) является метод усреднения результатов. Этот вопрос зачастую остается не освещенным, несмотря на значительное влияние его на результаты оценки.

В случае построения усредненной по множеству заданий той или иной множественной метрики можно рассмотреть две последовательности действий:

- сначала вычислить метрики по каждому запросу отдельно и затем их усреднить
- найти общее количество документов, относящихся к категориям таб.1 и уже на их основе вычислить искомую метрику

Первый способ вычисления принято называть макроусреднением (macroaverage), второй – микроусреднением (microaverage). Первый способ характерен для оценки задач поиска, в которых важен результат в среднем по запросу, независимо от мощности ответа на этот запрос. Второй же способ нашел большее применение в оценке классификации и фильтрации, где необходимо учитывать «размеры» запросов. В РОМИП'2003 при оценке результатов участников дорожки классификации применялись оба способа усреднения, а в дорожке поиска только микроусреднение.

3. Метрики на последовательностях документов

Метрики для списка документов, отсортированного по релевантности, учитывают не только факт наличия документа в списке найденных документов, но и его положение в этом списке. Такой вид метрик имеет смысл только в оценке систем поиска.

3.1. Точность на уровне n документов (precision(n))

Точность на уровне n документов определяется как количество релевантных документов среди первых n выданных документов, делённое на n [1].

Если система выдала более n документов, то эта величина равна точности системы на первых n документах результатов запроса. Если система выдала менее n документов, то точность на уровне n документов будет заведомо не выше точности системы.

Точность на уровне n документов характеризует способность системы выдавать релевантные документы в начале списка результатов. Например, если система выдает не более 10 документов на

первой странице, то $\text{precision}(10)$ отражает качество результатов системы, получаемых на первой странице.

Эта метрика имеет ряд недостатков. В частности, для различных запросов метрики $\text{precision}(n)$ могут быть несравнимы. Например, для «идеальной» системы, которая выдаёт только релевантные документы, $\text{precision}(100)=0.2$ для запроса, по которому существует 20 релевантных документов, и $\text{precision}(100)=0.3$ для запроса, по которому существует 30 релевантных документов. Несмотря на известные недостатки, точность на уровне является незаменимой метрикой современных систем поиска так как, в частности, позволяет оценить полезность первой страницы ответа системы для пользователя.

3.2. R-точность (R-precision)

R-точность равна точности на уровне n документов (п. 0) для n равного количеству релевантных документов для данного запроса [1].

Данная метрика особенно полезна в тех случаях, когда разных запросов наблюдается большая разница в количестве известных релевантных документов.

3.3. Средняя точность (average precision)

Средняя точность для данного запроса определяется следующим образом [1]: пусть для данного запроса имеется k релевантных документов.

Точность на уровне i -го релевантного документа $\text{prec_rel}(i)$ равна $\text{precision}(\text{pos}(i))$, если i -й релевантный документ находится в результатах запроса на позиции $\text{pos}(i)$. Если i -й релевантный документ не найден, то $\text{prec_rel}(i)=0$.

Средняя точность для данного запроса равна среднему значению величины $\text{prec_rel}(i)$ по всем k релевантным документам:

$$\text{AvgPrec} = \frac{1}{k} \sum_{i=1}^k \text{prec_rel}(i)$$

Отметим основные свойства метрики «средняя точность»:

- $\text{AvgPrec} \leq \text{recall}$
- если релевантные документы находятся только в начале списка результатов, то $\text{AvgPrec} = \text{recall}$
- если релевантные документы равномерно распределены по списку результатов, то $\text{AvgPrec} \approx \text{precision} \cdot \text{recall}$

- количество документов, ранжированных ниже последнего релевантного, не влияет на значение AvgPrec (отсекается «хвост»)

Средняя точность позволяет оценивать качество работы системы, учитывая приоритет высоко ранжированных документов перед документами, находящимися в конце списка. В отличие от метрик precision(n) и R-precision, средняя точность учитывает все найденные документы. В статье [2] отмечается, что метрика AvgPrec обладает высокой устойчивостью относительно вариаций оценки экспертов при вычислении средней оценки по множеству запросов.

3.4. «Ценность» ответа

Эта метрика (*ReciprocalRank*) позволяет оценить, сколько усилий требуется пользователю, чтобы найти первый ответ на свой вопрос, или какова вероятность того, что пользователь досмотрит результаты до позиции, где находится первый правильный ответ.

Формально, «ценность» ответа на конкретное задание вычисляется как:

$$\text{ReciprocalRank} = \text{rank}(\text{pos}_{rel}),$$

где pos_{rel} – это минимальная позиция, на которой находится релевантный ответ. Если правильных ответов в ответе нет, то «ценность» равна 0.

Функция $\text{rank}(\text{pos})$ обычно задается некоторой линейкой значений для нескольких первых позиций и считается равной 0 для всех остальных. Так, в дорожке QA конференции TREC [6] ненулевые ранги присваиваются только первым пяти ответам и линейка рангов выглядит как

$$\{1.0, 0.5, 0.33, 0.2, 0.1\}$$

(первая позиция – 1.0, вторая – 0.5, третья – 0.33, четвертая – 0.2, пятая – 0.1, все остальные - 0).

В дорожке вопросно-ответного поиска РОМИП также использовалась альтернативная линейка для 10 первых значений:

$$\{1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1\}$$

Наиболее часто, в литературе цитируются значения усредненной «ценности» ответов (*mean reciprocal rank*), представляющие собой среднее арифметическое рангов отдельных ответов.

4. 11-точечный график полноты/точности, измеренный по методике TREC (11-point matrix (TREC))

11-точечный график полноты/точности отражает изменение точности в зависимости от требований к полноте и дает более полную информацию, чем единая метрика в виде одной цифры [1, 5]. По оси абсцисс на графике откладываются значения полноты, по оси ординат – значение точности при условии, что рассматривается начальный отрезок результатов запроса, на котором достигается заданный уровень полноты.

Для запроса для которого известно n релевантных документов, полнота может принимать дискретные значения $0, 1/n, 2/n, \dots, 1$. Для того, чтобы можно было получать единый график полноты/точности для множества запросов

1. рассматриваются фиксированные значения полноты $0.0, 0.1, 0.2, \dots, 1.0$ (всего 11 значений);
2. используется специальная процедура интерполяции точности для данных фиксированных значений полноты;
3. для множества запросов производится усреднение точности для заданных уровней полноты.

Интерполированное значение точности равно максимальному значению точности при уровне полноты большем или равным заданному.

Сначала мы опишем процедуру вычисления интерполированных значений точности для данного запроса на примере, а затем дадим формальное описание этой процедуры.

Пример (из [5], см. рис. 1): пусть коллекция документов содержит 20 документов, 4 из которых релевантны запросу. Пусть система выдает в качестве результатов запроса все эти документы, ранжированные так, что релевантными являются первый, второй, четвертый и пятнадцатый. Для различных срезов результатов полнота принимает значения $0.25, 0.5, 0.75$ и 1.0 . В соответствии с правилом интерполяции, для значений полноты от 0 до 0.5 точность равна 1.0 , для значений полноты 0.6 и 0.7 точность равна 0.75 , для значений полноты $0.8, 0.9$ и 1.0 точность равна 0.27 ($4/15$).

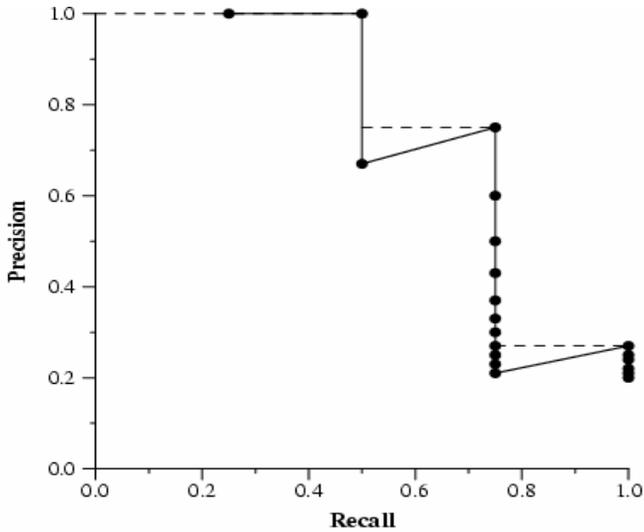


Рис. 2 Кривая полноты/точности для некоторого запроса. Точками обозначены значения полноты/точности для фиксированных срезов. Пунктирной линией – интерполированные значения.

Опишем процедуру построения 11-точечного графика более подробно. Для каждого значения полноты $r_i \in \{0.0, 0.1, 0.2, \dots, 1.0\}$ для каждого запроса q_j вычисляется значение интерполированной точности $p(r_i, q_j)$ следующим образом:

- если $r_i > \text{recall}(q_j)$, то $p(r_i, q_j) = 0$
- если $r_i \leq \text{recall}(q_j)$, то
 - $\text{pos}(r_i, q_j)$ равно минимальной длине списка результатов для запроса q_j , на которой достигается полнота r_i
 - $p(r_i, q_j) = \max_{n \geq \text{pos}(r_i, q_j)} (\text{precision}(n))$ (максимальная точность на начальном отрезке результатов длины $\text{pos}(r_i, q_j)$ или более).

Значение точности для множества запросов $\{q_j\}$ для фиксированного уровня полноты r_i равно среднему значению интерполированной точности для данного уровня полноты:

$$\text{Prec}(r_i) = \frac{1}{N} \sum_{j=1}^N p(r_i, q_j)$$

Литература

- [1] Program to evaluate TREC results using SMART evaluation procedures. Documentation. http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval/README
- [2] Buckley C., Voorhees E. Evaluating evaluation measure stability. In Proc. of the SIGIR'00, pp. 33-40, 2000.
- [3] C. J. van Rijsbergen. Information Retrieval. *Butterworth's and Co.*, London, U.K., 2 edition, 1979.
- [4] И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска. Программирование.28(4):226-242, 2002. <http://ir.apmath.spbu.ru/ru/papers.html>
- [5] The Twelfth Text Retrieval Conference (TREC 2003). Appendix 1. Common Evaluation Measures. <http://trec.nist.gov/pubs/trec12/appendices/measures.ps>
- [6] E. Voorhes. The TREC-8 Question Answering Track Report. In Proc. of the TREC-8, 1999.